Wikipedia Subcorpora Tool (WiST)

A tool for creating customized document collections for training unsupervised models of lexical semantics

Natalia Derbentseva
Peter Kwantes
DRDC – Toronto Research Centre

Simon Dennis
University of Newcastle

Benjamin Stone
Ohio State University

Defence Research and Development Canada

Scientific Report
DRDC-RDDC-2016-R100
June 2016

IMPORTANT INFORMATIVE STATEMENTS

This work was sponsored by the Influence Activities Task Force (IATF) and was conducted under project 15AH of the Command and Control thrust (5a).

Abstract

One of the most important advances in cognitive science over the past 20 years is the invention of computer models that can form semantic representations for words by analysing the patterns with which words are used in documents. Generally speaking, the models need to be 'trained' on tens of thousands of documents to form representations that are recognizable as the meaning or 'gist' of a term or document. Because the models derive meaning from words' usage across contexts/documents, the ways that words are used will drive the meaning. In this report, we describe the Wikipedia Subcorpora Tool (WiST), a tool for creating custom document corpora for the purpose of training models of lexical semantics. The tool is unique in that it allows the user to control the kinds of documents that comprise a corpus. For example, one might want to train a model to be an expert on medical topics, so the user can use the WiST to select a collection of medical documents on which to train the model. In this report, we detail the functionalities of the tool.

Significance to defence and security

Over the past decade, DRDC Toronto Research Centre has been exploring how computer models of lexical semantic can be embedded into software tools to support information search and analysis for practitioners in the intelligence community. The WiST tool is designed to improve the usefulness of such models in search and analysis tools by allowing semantic representations to be tailored specifically to particular domains.

Résumé

L’invention de modèles informatiques capables de créer les représentations sémantiques des mots à partir de leur distribution et de leurs occurrences dans les textes analysés constitue l’une des plus importantes avancées de la science cognitive au cours des 20 dernières années. En règle générale, des milliers de documents doivent servir à « entraîner » les modèles pour produire des représentations qui permettent de reconnaître la signification ou le « sens profond » d’un terme ou du contenu d’un document. Puisque les modèles interprètent le contexte ou le document à partir des mots employés, c’est la façon dont ils sont employés qui leur donne un sens. Dans le présent rapport, nous décrirons l’outil WiST (Wikipedia Subcorpora Tool) qui sert à créer et à personnaliser des corpus de documents dans le but d’entraîner des modèles de sémantique lexicale. Unique en son genre, WiST permet à l’utilisateur de décider des documents qui formeront un corpus. Ainsi, il pourra entraîner un modèle pour en faire un expert des sujets médicaux à partir d’un ensemble de documents médicaux sélectionnés à cette fin. Dans le présent rapport, nous décrivons en détail les fonctionnalités de l’outil.

Importance pour la défense et la sécurité

Au cours des dix dernières années, le Centre de recherche de Toronto de RDDC s’est penché sur la possibilité d’intégrer des modèles de sémantique lexicale à des outils logiciels pour répondre aux besoins touchant la recherche et l’analyse de l’information chez les praticiens de la communauté du renseignement. WiST est conçu pour améliorer l’utilité de tels modèles dans les outils de recherche et d’analyse en permettant l’adaptation de représentations sémantiques à des domaines précis.

Table of contents

This page intentionally left blank.

1 Introduction

Over the past twenty years several computational models have been developed to explain how the brain forms representations for the meanings of words from exposure to spoken and written language. Although formal models of semantic memory have existed since the 1960's, the new generation of models work very differently—a change owed largely to advances in computing power and its affordability. Early models (e.g., Collins & Quillian, 1968; Collins & Loftus, 1975) treated semantic memory as a network of connected concepts, represented as nodes. Activation of one node, say by presenting the model with the word, *dog*, would activate its corresponding node in the network, as well as all associated nodes like, *pet, leash, walk*, etc. The network of activated nodes therefore stood as the semantic representation of the concept, *dog.* Models such as the one just described are generally referred to as, *supervised models*, in that the connections among concepts in the network are hand-wired by the model builder. The new generation of models builds the associations among concepts without supervision. Generally speaking, new models work on the notion that words with related meaning occur in the same, or similar, contexts. Put another way, and in more specific terms, modern models of semantics build 'meaning' representations via a training phase during which they gather information about what terms tend to occur together (e.g., in the same document) in a large sample of documents, and from the co-occurrence information, infer what terms should occur together more generally in the language.

While in many cases building the model is straightforward, obtaining a corpus of documents on which to train the model can be a challenge. Most models perform at their best when tens of thousands of documents are used during training. For example, in Landauer and Dumais' (1997) seminal paper describing their model, Latent Semantic Analysis (LSA), they trained the system on 60,000 short documents extracted from an encyclopedia. We surmise that a tool for easily extracting training corpora for unsupervised semantics model would be of great use to theorists working in the domain. The first objective of the work reported here is to provide a tool that allows theorists to easily create training corpora for their models.

Another aspect of the work worth addressing is the role that context plays in the representation of meaning. Consider for a moment the banker who sails as a hobby. For her, the term *bank* has distinct meanings depending on where and when it is being used. At work, the term is used to describe the institution. At play, it is the part of a river that her sailboat must not hit. As we develop expertise in a domain, we develop an ability to selectively retrieve information relevant to the domain at the relative exclusion of more general knowledge (Ericsson & Delaney, 1998). In previous work, Terhaar and Kwantes (2010) simulated this ability to partition semantic knowledge by building models trained on documents relevant to a specific domain. There is currently no straightforward way to create a domain-specific training corpus for models like LSA. The second objective of the work reported here, is to give theorists the ability to arbitrarily define the domain from which the training documents are sampled.

We refer to the tool as the Wikipedia Subcorpora Tool (WiST). As suggested by its name, it uses Wikipedia as its primary document source. Documents are sampled from Wikipedia to create custom corpora of documents that can be used to train semantic models of language, such as LSA, mentioned above. Semantic associations between words are built from on their co-occurrences in the documents of the corpus. Based on the associations, LSA builds a mathematical representation of words' meanings, which can be used to compare the semantic similarity of texts without relying on exact word matches. To build an adequate semantic representation, however, LSA requires a large collection of short documents, usually in the thousands. Obtaining documents for a corpus can pose a logistical challenge. Corpus generation has usually been a labor intensive process, and the set of available corpora is relatively small.

Because the semantic models such as LSA rely on the co-occurrence of words in the corpus documents to construct their representations, the domain, from which the documents are selected for inclusion in the corpus, can have a significant impact on the resulting word representations. The most commonly used corpora for training various automatic and semi-automatic models of language are general collections of documents randomly extracted from a variety of subjects, for example Touchstone Applied Science Associates, Inc. (TASA) corpus and random selection of Wikipedia articles.

WiST was designed to automate the corpus generation and formatting process and to allow for creation of both general and topic-specific corpora to be used as training materials for semantic models. WiST works with a Wikipedia archive that must be extracted on the computer running the tool. The corpora that WiST generates are collections of Wikipedia articles that satisfy a user-provided search criteria. The remainder of this report describe WiST's functionality, corpus parameters set up, dependencies, and outputs.

2 The Wikipedia Subcorpora Tool (WiST)

Written in Python programming language, WiST is a tool that uses Wikipedia archive and Apache Lucene to generate custom corpora that can be used to train models of lexical semantics. Apache Lucene allows WiST to generate topic-specific corpora that contain articles on a given user-defined topic providing a greater homogeneity of content and semantic meaning of words in the corpus. WiST has flexible formatting parameters and it automates a number of corpus preparation activities such as it can remove punctuation marks and undesirable words (also known as 'stop words') from text and can format text such that the resulting corpus is ready to be used by a semantic model. This section describes how the tool works, its technical requirements, features and parameters, the tool's execution, output files and its potential applications.

2.1 How WiST works

To generate a corpus WiST selects a collection of articles from a Wikipedia archive. WiST relies on Apache Lucene,[1] an open source full text indexing and search engine, to retrieve articles from the archive to generate a topic-specific corpus. The user can specify the Lucene topic search query in the search query file (Section 2.4) and other corpus parameters and desired text formatting in the corpus configuration file (Section 2.5 and Annex B). Using the search query, WiST executes the Lucene search on the Wikipedia archive and uses search results to include articles in the corpus. Each article returned by Lucene search is formatted based on the formatting parameters specified in the configuration file (see Section 2.5 and Annex B) and then appended to

[1] https://lucene.apache.org/. WiST uses PyLucene extension of Apache Lucene.

the corpus file. After the required number of articles has been appended to the corpus file, the program removes words with frequency of occurrence that is lower than specified in the configuration file. The last operation performed on the text is the truncation of each article to a specified length, which is also indicated in the configuration file. On the output, WiST generates two files—a corpus file and a file that contains words that were removed from the corpus (see Section 2.7). The corpus file is a text file, which is ready for use by semantic models.

The user specifies all parameters necessary for generating a corpus with WiST in the configuration file (see Section 2.5 and Annex B), which is supplied as an option when executing WiST's main module (see Section 2.6). The list of parameters included in the configuration file also includes paths to three files:

- Lucene search query file;
- Stop words list file;
- Punctuation list file.

The user can modify these files based on their requirements, which provides greater flexibility for corpus generation.

Generating a corpus with WiST requires the following steps:

1. Ensure that all WiST dependencies are met on the computer that will be used to execute it, see Section 2.2;

2. Copy the WiST package to the computer on which it will be executed. See Section 2.3 for the list of the required and optional files for the WiST package;

3. Obtain a Wikipedia archive in .XML format. Instructions on where and how to download an English-language archive are available at:
https://en.wikipedia.org/wiki/Wikipedia:Database_download;

4. A Lucene index of the Wikipedia archive needs to be created, which can be done during the first run of the tool. See Annex B;

5. (Optional) Prepare the search query file. The search query file is required only when generating a topic-specific corpus. Instructions on search query file set up are in Section 2.4;

6. Prepare the corpus configuration file, see Section 2.5 and Annex B;

7. Execute the program to generate a corpus, see Section 2.6.

2.2 WiST dependencies

WiST is a Python module and requires the Python environment for its execution. Before WiST can be run, the following components must be installed:

- Python: *http://python.org/getit/*
- Java Development Kit (JDK): *http://www.oracle.com/technetwork/java/javase/downloads/index.html*
- Apache Ant: *http://ant.apache.org/bindownload.cgi*
- PyLucene: *http://www.apache.org/dyn/closer.cgi/lucene/pylucene/*
- A Wikipedia archive in the .XML format must be placed in the same directory as the WiST's main module. The Wikipedia archive can be downloaded from *http://download.wikimedia.org/enwiki/latest/enwiki-latest-pages-articles.xml.bz2* in the .BZ2 format, and it must be decompressed into its original .XML file—*wikipedia.xml* file.

WiST relies on Lucene index of the Wikipedia archive, which can be created during the first run of the tool. This index needs to be created only once for a given Wikipedia archive; however, this procedure needs to be repeated every time a new Wikipedia archive is extracted. Instructions on how to create a Lucene index are in the next section.

2.3 The WiST package

The main module of the tool is the file *WikipediaSubcorporaTool.py,* which is executed through a Python interpreter. For its execution, WiST also requires the following auxiliary files that contain functions for performing certain operations on the corpus. These files need to reside in the same directory as the main module:

- *default.cfg*—a default configuration file that is used by the tool if a custom configuration file is not supplied as an option at the time of execution (see Annex A);
- *CorpusCleaningTools.py*;
- *EntityClassify.py*;
- *WikiExtractor.py*;
- *IndexFiles.py*.

In addition to the files listed above, and depending on the specifications set in the configuration file, WiST may also require the following files for its execution:

- Configuration file with the .CFG extension is a customized configuration file, the file name is supplied with the '*-c'* option (see Section 2.6) at the time of execution. See Annex B on how to set up the configuration file;
- Search query file with the .TXT extension is required when creating a topic-specific corpus. The file contains key words and phrases that will be used to select Wikipedia articles that match the search criteria. See Section 2.4 on how to set up the search query file. The name of the search file is specified in the configuration file;
- Word stop-list file is a text file that contains all the words that will be removed from the corpus. The file must be formatted with a single word on each line;

- Punctuation stop-list file is a text file that contains all the punctuation characters that will be removed from the corpus text. The file must be formatted with a single character on each line.

The WiST package (excluding the software described in dependencies [Section 2.2] and Wikipedia archive) can be obtained by contacting the first author at natalia.derbentseva@drdc-rddc.gc.ca.

2.4 Search query file set up

WiST relies on Apache Lucene to retrieve articles from the Wikipedia archive. When a topic-specific corpus is desired the user defines the topic by specifying Lucene search query in the search query file. If the search query is left blank, a random set of articles will be retrieved.

The search query file is optional and should be used when generating a topic-specific corpus. The search query file contains the search keywords and phrases that Lucene search engine uses to retrieve relevant Wikipedia articles from the archive. If the number of articles that meet the search criteria is smaller than the specified size of the corpus, the corpus size will be limited to the number of available articles that meet the search criteria.

The search query file is a text file (with a .TXT extension), in which each line is a separate part of the query. When processing this file, WiST joins each line in this file with an “OR” operator. WiST will omit blank lines and lines that begin with a comment sign (“#”).

Lucene queries can contain AND, OR, NOT (-) and wild card operators, such as ‘*’ and ‘?’. Brackets () can be used to group parts of the query, and words can be grouped into phrases with double quotation marks.

An example of a query file is below:

```
# This is an example of a query file. The first two lines will not be included

# in the query because they begin with a comment sign (#)

(“computer programming” OR “programing language”) NOT (“objective C” OR Java*)

 “latent semantic analysis” AND “automated grading”

# this is the end of the query file. This line will also not be included in the query.
```

The search query file name to be used during the corpus generation must be specified in the configuration file with the *lucene_query_filename* = parameter (See Annex B).

2.5 Corpus parameters that can be set in WiST

The parameters for creating a corpus that can be set are:

- Corpus file name;

- Number of documents to be included in the corpus;
- Maximum document length in number of words. Documents will be truncated to this length;
- Minimum number of times a word must appear in the collection for it to be included in the corpus;
- Topic or search criteria: Search keywords for retrieving the documents from the archive. These must be saved in a separate file (see Section 2.4 on how to format this file) and this file's name (and path if stored in a different directory from the main module) must be included in the corpus configuration file. If no search query file name is provided in the configuration file, then a random set of documents will be retrieved;
- Tagging entities.

Corpus text preparation parameters (applied to all articles included in the corpus):

- Remove multiple white spaces;
- Remove formatting, e.g., paragraph and heading new lines;
- Remove new line characters from all documents. As a result, each document will be on a single line;
- Bringing all words to lower case;
- Remove single characters;
- Remove numbers;
- Remove stop words,[2] a file with a list of stop words must be provided;
- Remove punctuation, a file with a list of punctuation marks must be provided.

The above list is a set of actions that WiST can perform on the text. Each of these actions can be included or excluded depending on the desired result. All of these parameters are specified in the configuration file, which must be prepared prior to corpus generation. Annex B provides a detailed description of how to set up the configuration file and how to set all of the above parameters.

2.6 Executing the program and available options

WiST's main module, *WikipediaSubcorporaTool.py*, is a Python module and needs to be executed through a Python interpreter. The module takes the following options:

- --version show program's version number and exit;
- -h, --help show the help message and exit;

[2] Stop words are those words that the user wishes to remove for whatever reason from the articles. Often, words with the highest frequency of appearance in text are removed, because they do not allow discriminating among different contexts. Examples of stop words commonly removed from corpora include definite and indefinite articles, pronouns, prepositions, numbers, different versions of the verb "to be" etc. The user can create a custom stop word list to be applied to the corpus generated by WiST.

- -c CONFIG_FILENAME, --config=CONFIG_FILENAME Name of the configuration file (.cfg) to use, if this switch is omitted '*default.cfg'* (see Annex A) will be used.

Running the module with one of the first two options will display the requested information, i.e., the version number or the help message, and will exit without executing the rest of the code.

Running the *WikipediaSubcorporaTool.py* file with the last option and specified configuration file name (**or without any options**) will execute the code and will generate a corpus. All of the parameters for the new corpus to be generated are specified in the configuration file (See Section 2.5 and Annex B).

If *WikipediaSubcorporaTool.py* is run without specifying any option, then the module will generate a corpus based on the parameters set in the '*default.cfg*' file (see Annex A).

A topic-specific corpus of 10,000 articles can be generated with WiST in under 15 minutes.

2.7 WiST output files

WiST produces two output files:

- The corpus file that has the name specified with the *subcorpus_filename* parameter in the configuration file (see Annex B); and
- The file that contains all the words that were removed from the corpus because they did not meet the minimum occurrence criteria specified with the *term_minimum_occurrence* parameter (see Annex B). This file has the same name as the corpus file with an added extension of .REMOVED. This file contains the name of the corpus on the first line, date on the second line and each word is listed on a separate line.

The main output file is the corpus file, which is a plain text file that contains the specified number of Wikipedia articles, or as many articles as retrieved by Lucene with the given search query. Each article's text was processed based on the parameters specified in the configuration file (see Section 2.5 and Annex B). Usually the corpus file is formatted with each article on a single line with a blank line separating articles.

The size of the output corpus file in terms of the number of documents is specified by the user and can be as little or as large as is necessary. For example to train LSA models, a corpus of several tens of thousands of articles is desirable. However, if Lucene search returns fewer articles than was desired, the corpus will be limited to the number of articles that were returned by the search. If the number of returned articles is too few, the user can adjust the search criteria and repeat the process until the minimum number of articles is returned.

The actual file size of the resulting corpus depends on three criteria, all of which are defined by the user:

- The number of articles included;
- The number of stop words that are removed from the articles; and
- Truncation of each article.

For example, each 10,000 of articles takes up about 10MB, when each article is truncated to 300 words and when a stop word list of about 330 words is applied. Therefore, a 15,000-article corpus will be about 15MB, while a 50,000-article corpus will be roughly 50MB. If no truncation or stop word list is used, then the file size will be larger.

Two examples of a corpus file excerpt generated by WiST are provided in Annexes C and D. Annex C contains the first 20 articles from a corpus constructed from a random collection of articles; and Annex D contains the first 20 articles form a corpus on “military intelligence”. Both corpora were generated from the same Wikipedia archive and the same formatting criteria as described in Section 2.5 were applied to these two examples, truncating each article to 300 words.

All articles in Annex D are related to the topic of military intelligence and thus provide a more homogeneous context for word usage, whereas, articles in Annex C come from a large range of topics.

The content of a topic-specific corpus depends on the quality of the search query defined by the user, on the content of the archive and the quality of the Lucene search. Lucene is a well-known and widely used full text search engine.

2.8 WiST application

The relative ease with which WiST allows generating new corpora promotes testing and application of models of lexical semantics that require large corpora for training.

For example, Kwantes et al. (2014) used WiST to generate seven random and topic-specific corpora to assess participants’ personality traits from their essays. Kwantes et al. found that the agreement between essay’s LSA vectors and participants’ personality test scores improved when topic-specific corpora (i.e., trait-specific corpora in this case) were used over the randomly generated ones. This implies that LSA trained on topic-specific corpora could more accurately predict participants’ personality traits from their written essays.

Derbentseva et al. (2012) used a topic-specific corpus to train LSA in order to assess semantic similarity of concepts and propositions in the definitions of analytic integrity generated by groups of intelligence analysts. Because the goal of that work was to identify similar terms used by professionals in a specific domain, it was important to use a topic-specific corpus to ensure sensitivity of the analysis to that domain.

The on-going work at DRDC also investigates the application of custom topic-specific corpora to the analysis of Twitter content.

3 Conclusion

WiST is a useful tool for generating custom document corpora for the purpose of training models of lexical semantics. It is a flexible tool with many customisable parameters, and it relies on a Wikipedia archive, which provides a considerable pool of documents for corpus generation.

Currently, Wikipedia has over five million English articles, and new articles are added daily. Wikipedia archive can be periodically updated and indexed to ensure that the WiST's document pool remains current.

WiST can be used to generate general corpora from randomly selected documents; however its main advantage is the ability to create custom topic-specific corpora using a Lucene query. We believe that such a tool can facilitate the application of unsupervised semantic models, especially in context-specific domains.

This page intentionally left blank.

References

Collins, A.M. and Loftus, E.F. (1975). A spreading-activation theory of semantic processing. Psychological Review, Vol 82(6), 407–428.

Collins A.M. and Quillian, M.R. (1969). Retrieval time from semantic memory. Journal of Verbal Learning & Verbal Behavior, Vol 8(2), 240–247.

Derbentseva, N., Kwantes, P.J. and David Mandel (2012). Assessing diversity and similarity of conceptual understanding via semi-automated semantic analysis of concept maps. In proceedings of the 5th International Conference on Concept Mapping, Valletta, Malta. pp.168–175.

Kwantes, P.J., Derbentseva, N., Lam, Q., Vartanian, O. and Harvey H.C. Marmurek (2014). Assessing the Big Five personality traits with Latent Semantic Analysis. Unpublished manuscript.[3]

Landauer, T.K. and Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104(2), 211–240.

Terhaar, P. and Kwantes, P.J. (2010). Known Associations between Entities Impact Document Similarity Judgments: Implications for the Integration of Semantic Models into Intelligence Analysis Tools, *DRDC Toronto Technical Report*, TR 2010-071. Internally reviewed. 82 pages.

[3] This paper was accepted for publication in *Personality and Individual Differences* journal, but the authors had to withdraw it because DRDC could not negotiate an acceptable copyright agreement with the publisher.

This page intentionally left blank.

Annex A Default.cfg file

```
### Configuration file for wiki subcorpora generation

### There are three sections below:
### [Wikipedia] refers to the extraction of Wikipedia articles into
### text files from the Wikipedia xml file that can be downloaded here:
### http://download.wikimedia.org/enwiki/latest/enwiki-latest-pages-
articles.xml.bz2
### **Note: You will need to decompress the file from .bz2 into the
###         it's original xml
###
### [Lucene] refers to the configuration parameters to use with the
Lucene
### indexer. Lucene is used to index the Wikipedia articles and thus
retrieve
### articles that are related to boolean queries that you will construct
###
### [Subcorpus] parameters define the type or format of subcorpus that
you
### want to extract from the larger set of Wikipedia documents.
###
### ** IT IS REQUIRED THAT YOU SET THE 'True' or 'False' VALUE ON THE
'run' KEY
###    FOR EACH OF THESE THREE SECTIONS
###
### In all other cases, if Key/Values pairs that are removed, commented
out (#),
### or left equal blank (e.g., key = ) will not be included when
### wikipediaSubcorpora.py is run

[Wikipedia]
### Extract the text from the wikipedia articles
#run = False
run = True
### Filename of the wikipedia articles xml
#xml_filename = enwiki-latest-pages-articles.xml
xml_filename = smallWiki.xml
output_directory = .wiki_articles_text

[Lucene]
### Run the Lucene indexer on the text files
### in Wikipedia -> output_directory
run = True
### Location to store the lucene index
store_directory = .lucene_index
### Which Lucene Analyzer to use
analyzer = standard

[Subcorpus]
### Create a Subcorpora
run = True
```

```
### Filename for the subcorpora that you are going to create
subcorpus_filename =  mySubcorpus.cor
### The pattern for Lucene to match when retrieving the subcorpora,
### if commented out, left as an empty string or the key is not present
### then RANDOM documents will be retrieve by lucene
#lucene_query_filename = mySubcorpusLuceneQuery.txt
number_of_documents = 100
### Corpus Cleaning
### The cleaning value takes a list of operations seperated by '|' to
perform
### on the text. Operations are performed in the order that you specify,
and
### operations can be performed more than once.
###
### Possible cleaning operations are:
### * removeMultipleWhiteSpace - turns 'this  gap' into 'this gap'
### * removeFormatting - paragraph and heading newlines  and excess
whitespace
### * removeNewLines - remove newline characters from all articles
### * tagEntities - tag entities with underscores using python-NLTK,
###                 'Jimi Handrix' becomes '_Jimi_Hendrix_'.
###                  NOTE 1: NLTK is very slow!! Extracting named
entities
###                  from an individual document can take up to 10 secs,
so
###                  allow 24 hours to process a corpus of 10,000
documents
###                  NOTE 2: Perform this operation before lowering any
text
### * lowerText - lower case articles
### * removeSingleCharacters - remove any single character
### * removeNumbers - remove all numbers
### * removeSingleAlpha - remove all single letter a-zA-Z
### * removeStopWordsCaseSensitive - NOTE: word_stoplist must be set
below
### * removeStopWordsCaseInsensitive - NOTE: word_stoplist must be set
below
### * replacePunctuationWithSpace - NOTE: punctuation_stoplist must be
set below
### * replacePunctuationWithZeroSpace - NOTE: punctuation_stoplist must
be set below
###Slow:
cleaning_list =
tagEntities|lowerText|removeFormatting|replacePunctuationWithSpace|remov
eMultipleWhiteSpace|removeNumbers|removeSingleCharacters|removeStopWords
CaseInsensitive
###Quicker (without tagging named enitites):
#cleaning_list =
lowerText|removeFormatting|replacePunctuationWithSpace|removeMultipleWhi
teSpace|removeNumbers|removeSingleCharacters|removeStopWordsCaseInsensit
ive
### The word stoplist format must be a single term on each line
### on each line of the stoplist file (see stopList_Words.txt)
word_stoplist = stopList_Words.txt
```

```
### The punctuation stoplist format must be a single punctuation
### character on each line of the stoplist file (see
stopList_Punctuation.txt)
punctuation_stoplist = stopList_Punctuation.txt
### Minimum number of times a word must appear in a corpus for it to be
### include in the corpus. Removes very low frequency words
### removed terms will be recorded in [yourSubcorpusFilename].removed
term_minimum_occurrence = 2
### Truncates documents at max_document_word_length
max_document_word_length = 300
```

This page intentionally left blank.

Annex B Configuration file set up

The configuration file allows specifying parameters for generating a corpus. Configuration file is a text file with the extension .CFG, and it has three sections:

- [Wikipedia] – is the section that specifies parameters for the extraction of Wikipedia articles into text files from the Wikipedia XML file. This section must be run the first time a Wikipedia archive is used. In the subsequent runs this section can be disabled (with the RUN key set to FALSE);
- [Lucene] – is the section that specifies the configuration parameters for the Lucene indexer. Lucene is used to index the Wikipedia articles, which allows retrieving articles that are related to boolean queries supplied in the search query file. The Wikipedia archive must be indexed the first time it is used, and can be disabled in the subsequent runs (with the RUN key set to FALSE);
- [Subcorpus] – is the section that defines parameters for the type and format of the corpus to be created.

Each of these sections begins with the section heading—the name of the section written in square brackets on a new line, e.g.,

[Lucene]

The section heading is followed by a line with the RUN key, which must be set to either TRUE or FALSE to indicate whether the actions defined in the section will be executed or not.

Each section of the configuration file is described below.

B.1 [Wikipedia] section

The [Wikipedia] section instructs the program whether or not it needs to extract text from the Wikipedia articles stored in an XML Wikipedia archive and specifies the necessary parameters. The extraction of text is required for the Lucene indexer to process the articles, which enables the subsequent search and retrieval of the articles that match a given search criteria. There are three keys/parameters in this section:

- *run* = *True* (or *run* = *False*) indicates whether this section will be executed (= *True*) or not (= *False*);
- *xml_filename* = <name of the XML Wikipedia archive>. Indicates which XML Wikipedia archive to process for text extraction. E.g., *xml_filename* = *enwiki-latest-pages-articles.xml*;
- *output_directory* = <name of the directory to place the extracted text>. *E.g., output_directory* = *.wiki_articles_text*. The extracted text from the articles archives will be placed in the directory specified with this parameter. The Lucene indexer will analyse text stored in this directory specified with this parameter.

B.2 [Lucene] section

The [Lucene] section instructs the program whether or not the text extracted from a Wikipedia archive needs to be indexed and provides the necessary parameters. A Wikipedia archive needs to be indexed the first time it is used by WiST, and this action can be turned off for the subsequent runs. Note that turning off this section (with the "run" key set to "False") will not disable the ability to generate a topic-specific corpus. There are three keys/parameters in this section:

- *run = True* (or *run = False*) indicates whether this section will be executed (= *True*) or not (= *False*). If set to "True", the Lucene indexer will analyse text files in the *output_directory* specified in the [Wikipedia] section;
- *store_directory* = <name of the directory where to store the Lucene index>. E.g., *store_directory =.lucene_index*. This parameter indicates the location of the Lucene index. Lucene index is used to select articles for inclusion in the corpus;
- *analyzer* = <name of the analyser>. This parameter instructs the program which Lucene analyser to use. E.g., *analyzer = standard.*

B.3 [Subcorpus] section

All of the parameters for the new corpus are set in the [Subcorpus] section. There are **nine parameters** that can be set in this section, and they are described below.

- *run = True* (or *run = False*) indicates whether this section will be executed (= *True*) or not (= *False*). If it is set to *False*, the corpus will not be generated.
- subcorpus_filename = <name of the new corpus>. In this field the user specifies the name of the new corpus that will be created. If no path included in the file name, the file will be placed in the same directory where the main module resides; For example:

  subcorpus_filename = IntelligenceCorpus.cor
- *lucene_query_filename* = <name of the text file containing the search query for searching and selecting articles for the new corpus>. See Section 2.4 on how to set up a search query file. If this parameter is omitted (or commented out with the comment symbol "#") or left blank then random documents will be retrieved from the archive for the corpus. For example:

  lucene_query_filename = IntelligenceCorpusQuery.txt
- *number_of_documents* = <number of documents to be included in the corpus>. In this parameter the user specifies how many documents will be included in the new corpus. For example:

  number_of_documents = 15000
- *cleaning_list* = <list of cleaning operations separated by "|">. Text cleaning operations are performed after the articles are selected from the archive and all of the operations are applied to all of the articles selected for inclusion in the corpus. Operations are performed in the order they are listed and the same operation can be performed more than once. Possible cleaning operations are:

- *removeMultipleWhiteSpace*—removes multiple space characters leaving only one. E.g., turns ‘this gap’ into ‘this gap’;
- *removeFormatting*—removes paragraph and heading newlines and excess whitespace;
- *removeNewLines*—removes newline characters from all articles;
- *tagEntities*—tags entities with underscores using python Natural Language Tool Kit (NLTK). For example:

  ‘Jimi Hendrix’ becomes ‘_Jimi_Hendrix_’.

  Tagging entities must be performed before lowering case of the text. NLTK is very slow, therefore the processing can take a long time, e.g., 24 hours for 10,000 documents;
- *lowerText*—lowers case of all the text;
- *removeSingleCharacters*—removes any single character;
- *removeNumbers*—removes all numbers from all the text;
- *removeSingleAlpha*—removes all lower and upper case single letters a-z-A-Z;
- *removeStopWordsCaseSensitive*—removes all words (case sensitive) included in the word stop-list file specified with the *word_stoplist* parameter below. This operation should be performed before lowering case of the text;
- *removeStopWordsCaseInsensitive*—this operation is similar to the operation above (*removeStopWordsCaseSensitive*), with the difference that it disregards case of the words and, therefore can be performed before or after lowering case of the text. Similarly, this operation requires that a word stop-list file was set with the *word_stoplist* parameter below;
- *replacePunctuationWithSpace*—replaces punctuation characters listed in the punctuation stop-list file with a space character. The punctuation stop-list file must be specified in the *punctuation_stoplist* parameter below.
- *replacePunctuationWithZeroSpace*—similar to the above, this operation deletes punctuation characters listed in the punctuation stop-list file; however it only deletes the punctuation characters, it does not put a space character in their place. The punctuation stop list file must be specified in the *punctuation_stoplist* parameter below.

The following is an example of a *cleaning_list* string that includes tagging entities:

cleaning_list = tagEntities|lowerText|removeFormatting|replacePunctuationWithSpace|removeMultipleWhiteSpace|removeNumbers|removeSingleCharacters|removeStopWordsCaseInsensitive

The following is an example of a *cleaning_list* string **excluding** tagging entities. When *cleaning_list* is excluded, the routine executes much faster:

- *cleaning_list = lowerText|removeFormatting|replacePunctuationWithSpace|removeMultipleWhiteSpace|removeNumbers|removeSingleCharacters|removeStopWordsCaseInsensitive*

- *word_stoplist* = <name of the word stop-list file>. With this parameter the user indicates which word stop-list file will be applied to the articles selected for the corpus. The program will remove all words listed in the word stop-list file from all of the articles. The word stop-list file format must be a single term on each line of the file. The word stop-list filename (and path if different from the configuration file path) must be specified if either *removeStopWordsCaseSensitive* or *removeStopWordsCaseInsensitive* is included in the *cleaning_list* string. For example:

  word_stoplist = stopList_Words.txt

- *punctuation_stoplist* = < name of the punctuation stop-list file>. This parameter points to the file that contains the punctuation characters that will be removed from the corpus text. The format of the punctuation stop-list file must be a single punctuation character on each line of the file. The punctuation stop-list file must be specified if either *replacePunctuationWithSpace* or *replacePunctuationWithZeroSpace* is included in the *cleaning_list* string. For example:

  punctuation_stoplist = stopList_Punctuation.txt

- *term_minimum_occurrence* = <minimum number of times a word must appear in a corpus for it to be included in the corpus>. Words that do not meet the set minimum criteria will be removed from the corpus. For example:

  term_minimum_occurrence = 2

  This line indicates that words that occur only once in the corpus will be removed.

- max_document_word_length = <number of words in a document>. The value set in this parameter will be used to truncate each article to the specified length in number of words. For example:

  max_document_word_length = 300

  This line indicates that only the first 300 words in each document will be included in the corpus.

The *default.cfg* configuration file that is included with the package contains all possible parameters that can be included or adjusted during the corpus preparation process, and it serves as a starting point for customising the requirements for a new corpus. The *default.cfg* file relies on the comment sign (i.e., '#') to separate lines that will be included in the configuration from those that are omitted (because they are commented out). To ensure that all the necessary switches are included in the custom configuration file, the *default.cfg* file can be edited with a text editor and saved with a different name, which will be supplied with the "-c" option when executing the *WikipediaSubcorporaTool.py* file. The content of the *default.cfg* file is in Annex A.

Annex C An excerpt from a corpus generated from a random set of Wikipedia articles

```
fit flogging fit flogging leæther strip album released cleopatra records
penetrate satanic citizen bundles leæther strip releases prior solitary
confinement

kokopelli album kokopelli second album british band kosheen released uk moksha
recordings august album saw band lean rock music genre band member darren beale
states named album north american indian named kokopelli states spiritual
character used travel villages reservations spread fertility know make crops
grow suppose like witch doctor used music dance kind american history culture
sian reading guy thought link track listing songs written darren beale sian
evans substance track violence downloaded inserting kokopelli cd cd rom drive
going special area kosheen website available cd drive track automatically
downloaded free kbit s mp website longer exists kosheen site moksha recordings
violence released means violence currently available peer peer networks lyrics
violence web info remained printed uk album booklet website change

arkansas highway arkansas highway ar ark hwy series state highways run eastern
arkansas section highway state highway runs mississippi county begins
intersection ar maria heads east makes left turn heads north mile km turning
east continues east turning south ending intersection driver section highway
short state highway runs county begins access road mississippi river levee runs
west intersection rotan driver section highway state highway mississippi county
running osceola ar near victoria begins intersection north walnut street west
semmes avenue travels west intersecting southern terminus segment ar north ermen
lane northern terminus ar turns north west north crossing terminating ar outside
community victoria osceola spur highway short spur ar entirely osceola
mississippi county connects intersection ar north section ar section highway
state highway entirely mississippi county runs intersection ar north ar section
highway state highway runs mississippi county begins intersection ar poplar
corner heads north turning east mile km community buckeye makes left turn
missouri state line travels north entering community box elder ending state line
missouri supplemental routes continue north state line travel missouri section
highway state highway entirely clay county route runs north intersection ar near
community leonard intersection city rector

yes colombia yes colombia si colombia centrist political party colombia founded
noemí sanín dissident conservative party legislative elections march party won
small parties parliamentary representation

world world canadian broadcasting corporation flagship dinner hour radio
newscast airing monday friday local time cbc radio half hour program launched
saturdays sundays airs title world weekend maritime provinces world weekend airs
final hour live cross country checkup occupies time slot sundays atlantic time
zone rest canada world weekend airs local time simulcast cbc radio cbc radio
program airs radio march anchors program weekday anchor susan bonner september
weekend edition currently anchored marcia young past anchors reporters
associated program included alison smith joan donaldson maureen lorna jackson
barbara smith russ germain alannah campbell bob oxley bernie mcnamee dave

madhav mantri madhav mantri september indian cricketer played tests born nasik
maharashtra right handed opening batsman specialist wicket keeper represented
bombay captained bombay victory ranji trophy finals captained associated cement
company victory moin ud gold cup tournament mantri played test england india
```

toured england indian team playing tests pakistan test highest score bombay victory maharashtra semi final ranji trophy highest centuries match runs scored record class cricket mantri uncle indian cricket captain sunil gavaskar death lived hindu colony dadar mumbai oldest living indian test cricketer suffered heart attack hospitalized private clinic died following heart attack

union pines high school union pines high school year public high school located cameron north carolina opened school currently enrolls students public high schools moore county public school systems union pines sports teams cape fear valley conference includes schools harnett lee moore cumberland counties school fielded state championships wrestling tennis golf basketball individual state champions wrestling swimming tennis golf track field

david morris labour politician david morris january january welsh politician member european parliament mep chairman campaign nuclear disarmament cnd cymru peace activist morris born kidderminster adopted welsh family joined labour party age young man worked steel foundry llanelli south wales national service late exempted military service conscientious objector conditional working coal mines gained scholarship ruskin college oxford presbyterian minister morris anti nuclear campaigner opposing operation grapple britain tested nuclear weapons including hydrogen bombs pacific ocean atoll christmas island political career david morris served labour party councillor south wales elected european parliament mep boundary changes served representing south wales west area corresponding swansea neath port talbot bridgend late introduction list proportional representation british seats labour party introduced transitional selection process determine candidates european elections like internal labour party processes time labour london mayor selection welsh labour leadership election process determine order candidates party list elections controversial allegations undemocratic designed sideline left centre candidates morris morris like sitting welsh meps elected labour candidate members soon defunct constituency important process determine welsh labour candidates party list ranking morris placed low realistic chance elected withdrew candidate blamed outspoken opposition trident project retiring european parliament morris remained active welsh labour politics eventually benefited democratisation welsh labour party occurred rhodri morgan took leader elected represent south west wales area european constituency national executive committee welsh labour party served

seiji oko seiji oko seiji born february volleyball player japan member japan men national team won gold medal summer olympics silver medal summer olympics inductee volleyball hall fame holyoke massachusetts

terephthalic acid data page page provides supplementary chemical data terephthalic acid organic compound isomeric acids formula ch coh material safety data sheet handling chemical require notable safety precautions set forth material safety datasheet msds

omelek island omelek island pronounced kwajalein atoll republic marshall islands controlled united states military long term lease islands atoll ronald reagan ballistic missile defense test site geography island size geologically composed reef rock islands atoll created accumulation marine organism remnants corals mollusks history omelek long used united states small research rocket launches relative isolation south pacific government rocket launch occurred island equatorial proximity nearby radar tracking infrastructure attracted spacex orbital launch provider updated facilities island established primary launch location spacex began launching falcon rockets omelek falcon flight successful privately funded liquid propelled orbital launch vehicle launched omelek island september followed falcon launch july placing orbit omelek planned host launches upgraded falcon rocket spacex stopped development falcon launches focused large falcon launch manifest spacex tentatively planned upgrade launch site use falcon launch vehicle spacex launch manifest listed omelek kwajalein potential site falcon launches falcon overview document offered kwajalein launch option event spacex make upgrades necessary support falcon launches atoll reagan test site includes rocket launch sites islands kwajalein atoll wake island aur atoll government equatorial launch facility

multi chip module multi chip module mcm specialized electronic package multiple integrated circuits ics semiconductor dies discrete components packaged unifying substrate facilitating use single component larger ic mcm referred chip designs illustrating integrated nature overview multi chip modules come variety forms depending complexity development philosophies designers range using pre packaged ics small printed circuit board pcb meant mimic package footprint existing chip package fully custom chip packages integrating chip dies high density interconnection hdi substrate multi chip module packaging important facet modern electronic miniaturization micro electronic systems mcms classified according technology used create hdi high density interconnection substrate chip stack mcms relatively new development mcm technology called chip stack package certain ics memories particular similar identical pinouts used multiple times systems carefully designed substrate allow dies stacked vertical configuration making resultant mcm footprint smaller albeit cost thicker taller chip area premium miniature electronics designs chip stack attractive option applications cell phones personal digital assistants pdas thinning process dies stacked create high capacity sd memory card

institute mathematics physics mechanics institute mathematics physics mechanics abbreviation leading research institution area mathematics theoretical science slovenia includes researchers university ljubljana university maribor university founded

fred borch colonel frederic borch born career united states army attorney master national security studies served chief prosecutor guantanamo military commissions resigned commission august prosecutors complained rigged providing process defendants replaced robert swann worked time civilian consultant prosecution teams guantanamo military commissions hired position archive historian judge advocate general corps awarded fulbright fellowship serve visiting professor university leiden teaching issues terrorism counter terrorism education borch earned history davidson college commissioned army studied law university north carolina degree university brussels ll international comparative law magna cum laude military career legal assistant fort benning army infantry school th infantry regiment borch spent years defense counsel army trial defense service kaiserslautern germany borch enrolled year judge advocate general school charlottesville virginia received degree military law assigned fort bragg xviii airborne corps serving civilian assistant district attorney north carolina borch began year term professor criminal law jag school specialising fourth amendment application following position studied command general staff college fort leavenworth assigned job joint service committee military justice jag office pentagon drafted legislation related uniform code military justice proposed changes manual courts martial borch oversaw successful prosecution drill sergeants accused sexual misconduct aberdeen proving ground promoted deputy chief army s government appellate division following year staff judge advocate fort gordon army signal center attended naval war college newport rhode island graduated class receiving masters degree national security studies took position professor international law focusing counter terrorism guantanamo bay military commission responding united states supreme court decision rasul bush detainees right challenge detention impartial tribunal department defense set combatant status review tribunals administrative review boards military commissions try defendants charged war crimes borch appointed chief prosecutor military commission worked prepare trials starting alleged corruption guantanamo hearings august prosecutors capt john carr maj robert preston wrote borch told presiding officers chosen sure convict reportedly said evidence shawnee taveras shawnee taveras dominican american singer songwriter specializing merengue genre media personality living providence rhode island taveras member telemundo providence cast released singles received critical acclaim northeast dominican republic single sé como duele reached hit list santo domingo late shawnee appeared prime time television radio shows dominican republic including el jochy santos te estoy la belleza es mia mia cepeda el zol la mañana el mismo golpe received award artistic excellence dominican republic ministry youth taveras born santo domingo family moved rhode island young started music career providence participating el festival la voz shortly decided pursue dreams music performing festivals concerts region performed dominican republic musical influences include milly quezada shakira juan luis guerra shawnee currently psychology major roger williams university

kamada surname written kamata vice admiral imperial japanese navy saw service pacific theatre world war ii biography kamada native ehime prefecture shikoku island japan graduated th class imperial japanese naval academy ranked th class classmates included future admirals takeo takagi hara shigeyoshi miwa sadamichi served midshipman duty cruisers sub lieutenant battleship cruiser battlecruiser destroyer promoted lieutenant serving battleship assigned survey ships musashi yamato chief gunnery officer battleship february promotion lieutenant commander december served cruisers receiving command destroyer november promotion commander december kamada served executive officer battleship november promoted captain november captain cruiser subsequently commanded cruisers izumo appointed imperial japanese navy general staff october stationed japanese occupied hainan island kamada promoted rear admiral october served staff commanded forces japanese th fleet new guinea october december august kamada took command japanese naval forces designated nd naval special base force based balikpapan borneo making military governor dutch borneo kamada forces subsequently involved borneo campaign promoted vice admiral kamada surrendered forces australian major general edward james milford aboard september surrender japan dutch military court pontianak convicted war crimes executions west borneo natives ill treatment dutch pows held flores island kamada sentenced death executed october

charles carrington charles carrington leading british publisher erotica late th early th century europe born paul harry ferdinando bethnal green england november moved london paris published sold books rue faubourg montmartre rue short period moved activities brussels carrington published works classical literature including english translation aristophanes comedies books famous authors oscar wilde anatole france order hide undercover erotica publications veil legitimacy books featured erotic art martin van published french series la flagellation travers le monde mainly english flagellation identifying english predilection carrington blind result syphilis years life spent poverty mistress stole valuable collection rare books placed lunatic asylum died ivry sur seine france

battle mhlatuze river battle mhlatuze river battle fought zulu ndwandwe tribes following zulu civil war ndwandwe hierarchy set asunder battle largely scattered population response history shaka attacked warriors led river battle zulu people prevailed battle led military commander shaka battle hill shaka superior tactics led people victory ndwandwe attack came waited half river effectively splitting attackers separate groups allowed zulu victory

hairpin lace hairpin lace lace making technique crochet hook small hairpin lace loom used loom consisting parallel metal rods held removable bars historically metal shaped hairpin used originates hairpin lace formed wrapping yarn prongs hairpin lace loom form loops held row crochet stitched worked center called spine resulting piece lace worked length desired removing bar hairpin slipping loops end strips produced process joined create airy lightweight fabric various types yarns threads used achieve different color texture design effects examples items hairpin lace include scarves shawls hats baby blankets afghans clothing hairpin lace added sewn knitted crocheted works decorative accent

moran town moran town census town dibrugarh district indian state assam moran important industrial town india major oil field major tea producing area geography moran located average elevation demographics india census moran town population males constitute population females moran town average literacy rate higher national average male literacy female literacy moran town population years age government moran dibrugarh lok sabha constituency

This page intentionally left blank.

Annex D An excerpt from a corpus generated on the ‘military intelligence’ topic

Search query used to generate this corpus was:
(“intelligence analysis” OR “military intelligence” OR “intelligence assessment” OR “defence intelligence”) NOT (“intellectual competence” OR “intellect” OR “IQ test”)

```
military intelligence board military intelligence board mib serves senior level
board coordination intelligence assets support military operations globally
board chaired defense intelligence agency seeks consensus commands agencies
services forum discuss intelligence issues related military combat operations
mib meets daily address coordinate intelligence analysis assets collection
systems personnel

director defense intelligence agency director defense intelligence agency star
general admiral nomination president confirmation senate serves nation highest
ranking military intelligence officer primary intelligence adviser secretary
defense chairman joint chiefs staff answers director national intelligence
civilian secretary defense intelligence director commander joint functional
component command intelligence surveillance reconnaissance subordinate command
united states strategic command additionally chairs military intelligence board
coordinates activities entire defense intelligence community

james williams james williams born march retired united states army lieutenant
general williams served director defense intelligence agency inductee military
intelligence hall fame chairman board directors national military intelligence
association early life education williams born paterson new jersey march youth
williams paid visit military academy years later began federal service volunteer
aircraft spotter nd anti aircraft region youth active sports playing baseball
running track swimming member nj group ii state basketball championship team
garnered group ii state honors avid hiker skier williams graduated united states
military academy bachelor science degree engineering initially commissioned
second lieutenant air defense artillery received master arts degree latin
american studies university new mexico military education includes completion
air defense basic officers course united states army intelligence school united
states army command general staff college defense intelligence school national
war college career williams began career air defense artillery assignments tour
field command subsequent assignments intelligence field assignments detachment
fort amador canal zone st counterintelligence corps detachment fort brooke
puerto rico assigned fort leavenworth kansas project intelligence officer
tactical aerial reconnaissance surveillance tars served assistant army attaché
caracas venezuela commanded st military intelligence battalion provisional th
military intelligence group united states army vietnam supporting iii marine
amphibious force iii maf served washington dc remaining washington williams
named military affairs bureau inter american affairs state department assumed
command counterintelligence supreme headquarters allied powers europe return
united states assigned defense intelligence agency chief missile forces
strategic arms limitation branch soviet warsaw pact division later served deputy
director estimates prior return defense intelligence agency director williams
served deputy chief staff intelligence united states army europe leading team
dod analysts provide strategic early warning contingency planning martial law
soviet dominated poland vice warsaw pact intervention september appointed
director dia williams culminated years service year tour director defense
intelligence agency dia senior
```

intelligence gathering network intelligence gathering network information particular entity collected benefit use inter related source information gathered military intelligence government intelligence commercial intelligence network intelligence assessment employs intelligence analysis refine information foreign embassies subscribe newspapers tabs news channels host countries information doesn classified considered useful intelligence called osint open source intelligence increasing quantity utility ascendancy digital media researchers employed dig archives check facts important form intelligence called signals intelligence attempts intercept electronic communications signals sent parties working hostile potentially hostile entity neutral friendly parties discussing entity established intelligence agencies networks usually follow linear distributed structure agent handler directs activities number persons sources order obtain necessary facts target intelligence gathering operation main humint agent types used infiltration penetration agents infiltration agent enters target operation outside suitable pretext suspected espionage penetration agent place target area recruited handler means mice principle information gathered processed analysts turned intelligence product information conveyed nodes network variety secure clandestine means physical electronic

joint intelligence organisation united kingdom joint intelligence organisation british intelligence agency responsible intelligence assessment development uk intelligence community s analytical capability headed jon day permanent secretary level civil servant organisation supports work joint intelligence committee tasked directing secret intelligence service security service gchq national security council providing intelligence assessments ministers senior officials intelligence assessment primary function organisation provide assessments situations issues current concern warnings threats british interests identifying monitoring countries risk instability consisting intelligence analysts wide range departments disciplines assessments staff draws range intelligence primarily british intelligence agencies diplomatic reporting open source material joint intelligence committee agrees assessments circulated ministers senior officials professional head intelligence analysis head jio professional head intelligence analysis advises gaps duplication analyst training recruitment analysts career structures interchange opportunities order improve uk intelligence community s analytical capability professional head intelligence analysis carries development analytical methodology training uk intelligence community intelligence analysts

history espionage espionage intelligence assessment existed ancient times pre modern espionage early strategists sun zi stressed need military intelligence modern times modern age came concept professional police organizations police state geopolitics new intelligence methods arrived imagery intelligence signals intelligence cryptanalysis spy satellites

defense intelligence agency headquarters defense intelligence agency headquarters dia hq main operating center defense intelligence agency located premises joint base anacostia bolling washington dc overview dia headquarters originally called defense intelligence analysis center diac fully completed designed smithgroupjjr consolidate dia activities washington dc area agency opened headquarters expansion designed smithgroupjjr allowed dia personnel serve roof simultaneously housed office director national intelligence dni facility opened liberty crossing mclean va dia hq headquarters national intelligence university located united states strategic command s joint functional component command intelligence surveillance reconnaissance jfcc isr dia hq includes patriot memorial commemorates defense intelligence agency employees died service agency united states additionally facility houses memorial honoring seven employees died attacks september pentagon torch bearers wall recognizes employees exceptional contributions agency s mission currently approximately dia workforce serves headquarters patrick lang walter patrick pat lang jr born commentator middle east retired army officer private intelligence analyst author leaving uniformed military service colonel held high level posts military intelligence civilian led intelligence analysis middle east south asia defense department world wide humint activities high level equivalent rank lieutenant general background lang graduated virginia military institute ba english university utah ma middle east studies member phi kappa phi lang member equestrian order holy sepulchre roman catholic chivalric order holds rank knight commander personal life married marguerite lessard reside alexandria virginia uncle john lang served world wars interwar period canadian military forces received military honors actions bravery united kingdom united states japan military service serving army lang graduated army war college army command general staff college armed forces staff college decorated veteran united states overseas conflicts vietnam war served special forces military intelligence trained educated specialist middle east served region years professor arabic united states military academy twice selected best classroom teacher year defense intelligence agency defense intelligence officer dio middle east south asia counter terrorism later director defense humint service dia member defense senior executive service participated drafting national intelligence estimates military attachés worldwide reported period briefed president george bush white house operation desert storm head intelligence analysis middle east seven years institution head middle east south asia analysis dia counter terrorism seven years service dia lang received presidential rank award distinguished executive post retirement activities leaving government service joined veteran intelligence professionals sanity left group policy differences period prior iraq war registered department justice foreign agents registration act work behalf lebanese politician industrialist promoted peace process vocational training building trades english french language instruction extending microcredit registered advice counsel deregistered continuing work peace process participated work harry frank guggenheim foundation example foundation sponsors individuals scholarly research violence aggression dominance lang

joint analysis center joint intelligence operations center europe jioceur analytic center jac known joint analysis center joint intelligence center serving focal point military intelligence united states european command located raf molesworth cambridgeshire uk managed defense intelligence agency area responsibility includes countries europe middle east

joint intelligence operations center europe analytic center joint intelligence operations center europe jioceur analytic center jac known joint analysis center joint intelligence center serving focal point military intelligence united states european command located raf molesworth cambridgeshire uk jioceur administered defense intelligence agency area responsibility includes countries europe middle east

intelligence agency intelligence agency government agency responsible collection analysis exploitation information intelligence support law enforcement national security defence foreign policy objectives means information gathering overt covert include espionage communication interception cryptanalysis cooperation institutions evaluation public sources assembly propagation information known intelligence analysis intelligence assessment intelligence agencies provide following services national governments distinction security intelligence foreign intelligence security intelligence pertains domestic threats terrorism espionage foreign intelligence involves information collection relating political economic activities foreign states agencies involved assassination arms trafficking coups état placement misinformation propaganda covert operations order support governments interests

eu intelligence analysis centre eu intcen eu intelligence analysis centre eu intcen intelligence body european union eu january eu intcen european external action service eeas authority eu high representative mission eu intcen mission provide intelligence analysis early warning situational awareness high representative federica mogherini european external action service various eu decision making bodies fields common security foreign policy common security defence policy counter terrorism eu member states eu intcen monitoring assessing international events focusing particularly sensitive geographical areas terrorism proliferation weapons mass destruction global threats history eu intcen roots european security defence policy group analysts working open source intelligence supervision high representative javier solana called joint situation centre wake terrorist attacks new york washington september solana decided use existing joint situation centre start producing intelligence based classified assessments request solana council european union agreed june establish sitcen counter terrorist cell cell tasked produce counter terrorist intelligence analyses support member states security services sitcen generally used eu situation centre officially renamed european union intelligence analysis centre eu intcen organisation total number eu intcen staff close single intelligence analysis capacity siac eu intcen single intelligence analysis capacity siac combines civilian intelligence eu intcen military intelligence eums intelligence directorate framework siac civilian military contributions used produce source intelligence assessments eu intcen eums intelligence directorate main clients european union satellite centre provides satellite imagery analysis

joint intelligence center joint intelligence center jic focal point military intelligence gathered different intelligence agencies administered defense intelligence agency intelligence center joint force headquarters joint intelligence center responsible providing producing intelligence required support joint force commander staff components task forces elements national intelligence community joint intelligence centers united states central command tampa florida united states pacific command hawaii europe joint analysis center serves jic united states european command

defence staff intelligence division defence staff intelligence division military intelligence agency malaysia armed forces role said equivalent defense intelligence agency dsid headed army lieutenant general consists tri services military branch army intelligence naval intelligence air force intelligence head dsid known director general reporting directly chief armed forces reports minister defence national security division current director lt gen dato paduka abdul hadi haji hussin

edmund thompson edmund thompson united states army general officer military career july august thompson brigadier general commanding general army intelligence agency august november thompson major general served assistant chief staff intelligence department army headquarters deputy director management operations defense intelligence agency general thompson member military intelligence hall fame

dennis nagy dennis mark nagy born acting director defense intelligence agency september november background nagy hungarian ancestry attended air force academy graduated bachelor science degree international relations commission air force served pilot air force attended graduate school georgetown university international relations began dia career july intelligence analyst assignment permanent assignment newly formed directorate estimates succession progressively responsible assignments nagy focused soviet strategic nuclear space forces policy doctrinal issues principal drafter numerous departmental national estimates service culminated selected twice director central intelligence s national intelligence officer strategic programs manager annual national intelligence estimate soviet strategic nuclear forces late nagy dia executive selected deputy vice director overall management program development estimative basic scientific technical intelligence production extended periods nagy served acting vice director soviet military power nagy personally directed development issue soviet military power dod s annual publication soviet military policies forces september nagy charter member defense intelligence senior executive service nagy appointed position assistant deputy director research capacity held position chief directorate research db dia s largest single military intelligence production organization served general defense intelligence program gdip functional manager general military intelligence chairman council defense intelligence producers military targeting committee nagy appointed position executive director dia elevating agency s command element ranking agency s senior civilian nagy appointed deputy director acting director september director nagy appointed acting director interim period september november civilian placed position acting director provided continuity critical time decrements agency resources caused reconsideration managerial issues review traditional threat priorities defense intelligence community served lieutenant general james clapper jr usaf assumed directorship

patrick hughes patrick hughes born september retired united states army officer served th director defense intelligence agency previously director intelligence dia joint staff office chairman joint chiefs staff director intelligence united states central command commanding general united states army intelligence agency army joined united states department homeland security assistant secretary information analysis intelligence departed dhs government service march early life education hughes born september great falls montana shortly birth family moved small town manhattan montana gallatin valley near bozeman raised schooled formative years hughes active sports school activities held variety jobs young age spent summers riverton wyoming jackson hole wyoming father worked lived graduated attended montana state college later designated university brigham young university provo utah joining army january following initial enlistment hughes returned montana state university january pursue college education degree hughes commissioned army rotc program montana state university bozeman montana june earned bachelor arts degree business earned master arts business management central michigan university concurrent graduation army command general staff college hughes attended school advanced military studies sams advanced operational studies fellow aosf lieu attendance war college received honorary doctorates montana state university business national defense intelligence college military intelligence military education training includes infantry officer basic course iobc ra fort benning georgia military assistance training advisor mata course fort bragg counterintelligence research officer course fort holabird maryland military assistance security adviser masa course fort bragg united states army intelligence center training military intelligence officers advanced course fort huachuca completed army basic training army medical corpsman training combat medic single engine pilot training basic airborne school jumpmaster training jungle warfare school operations course trained vietnamese language conjunction mata masa training john kennedy special warfare center school fort bragg later korean language defense language institute monterey california completed electronic warfare cryptology officer familiarization course advanced military studies program war

marine corps intelligence activity marine corps intelligence activity mcia field activity headquarters marine corps member defense intelligence agency united states intelligence community mcia describes vital military intelligence corporate enterprise functions collegial effective manner service agencies joint intelligence centers joint chiefs staff unified commands marine corps intelligence activity mission provide intelligence services marine corps intelligence community services based expeditionary mission profiles littoral areas supports development service doctrine force structure training education acquisition mcia determines missions corps needs carry need trained mission mcia partnership office naval intelligence office coast guard intelligence national maritime intelligence integration office marine corps base quantico quantico virginia united states intelligence community united states intelligence community federation separate united states government agencies work separately conduct intelligence activities considered necessary conduct foreign relations national security united states member organizations include intelligence agencies military intelligence civilian intelligence analysis offices federal executive departments headed director national intelligence dni reports president united states varied responsibilities members community collect produce foreign domestic intelligence contribute military planning perform espionage established executive order signed december president ronald reagan washington post reported government organizations private companies locations united states working counterterrorism homeland security intelligence intelligence community includes people holding secret clearances according study office director national intelligence private contractors make workforce intelligence community cost equivalent personnel budgets etymology term intelligence community used lt gen walter bedell smith tenure director central intelligence history collection analysis production sensitive information support national security leaders including policymakers military commanders members congress safeguarding processes information counterintelligence activities execution covert operations approved president ic strives provide valuable insight important issues gathering raw intelligence analyzing data context producing timely relevant products customers levels national security from war fighter ground president washington members ic consists members called elements offices bureaus federal executive departments ic led director national intelligence programs definitions nip mip overlap address military intelligence assignment department defense intelligence activities nip mip proves problematic organizational structure leadership overall organization ic primarily governed national security act amended statutory organizational relationships substantially revised intelligence reform terrorism prevention act irtpa amendments national security act ic characterizes federation member elements overall structure better characterized confederation lack defined unified leadership governance structure prior director central intelligence dci head ic addition director cia major criticism arrangement dci little actual authority budgetary authorities ic agencies limited influence operations dni authority direct control element ic staff office dni dni authority hire personnel ic staff member elements executive branch directed controlled respective

secretary defense intelligence secretary intelligence usd high ranking civilian position office secretary defense osd department defense acts principal civilian advisor deputy secretary deputy secretary defense matters relating military intelligence secretary appointed civilian life president confirmed senate serve pleasure president overview office secretary defense intelligence ousd principal staff element department defense regarding intelligence counterintelligence security sensitive activities intelligence related matters secretary defense representative usd exercises oversight defense intelligence agency dia national geospatial intelligence agency nga national reconnaissance office nro national security agency nsa addition secretary dual hatted serving director defense intelligence office director national intelligence rank secretary usd level iii position executive schedule january annual rate pay level iii history position secretary defense intelligence created national defense authorization act fiscal year aftermath september terror attacks better coordinate department wide intelligence activities second line succession secretary defense deputy secretary defense executive order president george bush december created legislation described taking precedence department secretary personnel readiness november department defense directive secretary rumsfeld stated secretary shall serve secretary primary representative office director national intelligence stated secretary shall provide policy oversight training career development personnel department defense counterterrorism intelligence security components secretary duty finding candidates nominated serve directors defense intelligence agency national geospatial intelligence agency national reconnaissance office

```
national security agency overseeing performance usd dual hatted position
director defense intelligence acting primary military intelligence advisor dni
additional position follows memorandum agreement secretary defense robert gates
director national intelligence john michael mcconnell create position office
secretary secretary leads office secretary defense intelligence ousd unit office
secretary defense ousd exercises planning policy strategic oversight department
defense intelligence counterintelligence security matters ousd serves primary
representative defense department director national intelligence members united
states intelligence community budget totals annual budget usd contained office
secretary defense osd budget defense wide operation maintenance account
```

This page intentionally left blank.

List of symbols/abbreviations/acronyms/initialisms

DND	Department of National Defence
DRDC	Defence Research and Development Canada
DSTKIM	Director Science and Technology Knowledge and Information Management
JDK	Java Development Kit
LSA	Latent Semantic Analysis
NLTK	Natural Language Tool Kit
R&D	Research & Development
TASA	Touchstone Applied Science Associates
WiST	Wikipedia Subcorpora Tool

This page intentionally left blank.

DOCUMENT CONTROL DATA

(Security markings for the title, abstract and indexing annotation must be entered when the document is Classified or Designated)

1. ORIGINATOR (The name and address of the organization preparing the document. Organizations for whom the document was prepared, e.g., Centre sponsoring a contractor's report, or tasking agency, are entered in Section 8.)

DRDC – Toronto Research Centre
Defence Research and Development Canada
1133 Sheppard Avenue West
P.O. Box 2000
Toronto, Ontario M3M 3B9
Canada

2a. SECURITY MARKING (Overall security marking of the document including special supplemental markings if applicable.)

UNCLASSIFIED

2b. CONTROLLED GOODS

(NON-CONTROLLED GOODS)
DMC A
REVIEW: GCEC DECEMBER 2013

3. TITLE (The complete document title as indicated on the title page. Its classification should be indicated by the appropriate abbreviation (S, C or U) in parentheses after the title.)

Wikipedia Subcorpora Tool (WiST) : A tool for creating customized document collections for training unsupervised models of lexical semantics

4. AUTHORS (last name, followed by initials – ranks, titles, etc., not to be used)

Derbentseva, N.; Kwantes, P.; Dennis, S.; Stone, B.

5. DATE OF PUBLICATION (Month and year of publication of document.)

June 2016

6a. NO. OF PAGES (Total containing information, including Annexes, Appendices, etc.)

44

6b. NO. OF REFS (Total cited in document.)

6

7. DESCRIPTIVE NOTES (The category of the document, e.g., technical report, technical note or memorandum. If appropriate, enter the type of report, e.g., interim, progress, summary, annual or final. Give the inclusive dates when a specific reporting period is covered.)

Scientific Report

8. SPONSORING ACTIVITY (The name of the department project office or laboratory sponsoring the research and development – include address.)

DRDC – Toronto Research Centre
Defence Research and Development Canada
1133 Sheppard Avenue West
P.O. Box 2000
Toronto, Ontario M3M 3B9
Canada

9a. PROJECT OR GRANT NO. (If appropriate, the applicable research and development project or grant number under which the document was written. Please specify whether project or grant.)

15ah

9b. CONTRACT NO. (If appropriate, the applicable number under which the document was written.)

10a. ORIGINATOR'S DOCUMENT NUMBER (The official document number by which the document is identified by the originating activity. This number must be unique to this document.)

DRDC-RDDC-2016-R100

10b. OTHER DOCUMENT NO(s). (Any other numbers which may be assigned this document either by the originator or by the sponsor.)

11. DOCUMENT AVAILABILITY (Any limitations on further dissemination of the document, other than those imposed by security classification.)

Unlimited

12. DOCUMENT ANNOUNCEMENT (Any limitation to the bibliographic announcement of this document. This will normally correspond to the Document Availability (11). However, where further distribution (beyond the audience specified in (11) is possible, a wider announcement audience may be selected.))

Unlimited

13. ABSTRACT (A brief and factual summary of the document. It may also appear elsewhere in the body of the document itself. It is highly desirable that the abstract of classified documents be unclassified. Each paragraph of the abstract shall begin with an indication of the security classification of the information in the paragraph (unless the document itself is unclassified) represented as (S), (C), (R), or (U). It is not necessary to include here abstracts in both official languages unless the text is bilingual.)

One of the most important advances in cognitive science over the past 20 years is the invention of computer models that can form semantic representations for words by analysing the patterns with which words are used in documents. Generally speaking, the models need to be 'trained' on tens of thousands of documents to form representations that are recognizable as the meaning or 'gist' of a term or document. Because the models derive meaning from words' usage across contexts/documents, the ways that words are used will drive the meaning. In this report, we describe the Wikipedia Subcorpora Tool (WiST), a tool for creating custom document corpora for the purpose of training models of lexical semantics. The tool is unique in that it allows the user to control the kinds of documents that comprise a corpus. For example, one might want to train a model to be an expert on medical topics, so the user can use the WiST to select a collection of medical documents on which to train the model. In this report, we detail the functionalities of the tool.

L'invention de modèles informatiques capables de créer les représentations sémantiques des mots à partir de leur distribution et de leurs occurrences dans les textes analysés constitue l'une des plus importantes avancées de la science cognitive au cours des 20 dernières années. En règle générale, des milliers de documents doivent servir à « entraîner » les modèles pour produire des représentations qui permettent de reconnaître la signification ou le « sens profond » d'un terme ou du contenu d'un document. Puisque les modèles interprètent le contexte ou le document à partir des mots employés, c'est la façon dont ils sont employés qui leur donne un sens. Dans le présent rapport, nous décrirons l'outil WiST (Wikipedia Subcorpora Tool) qui sert à créer et à personnaliser des corpus de documents dans le but d'entraîner des modèles de sémantique lexicale. Unique en son genre, WiST permet à l'utilisateur de décider des documents qui formeront un corpus. Ainsi, il pourra entraîner un modèle pour en faire un expert des sujets médicaux à partir d'un ensemble de documents médicaux sélectionnés à cette fin. Dans le présent rapport, nous décrivons en détail les fonctionnalités de l'outil.

14. KEYWORDS, DESCRIPTORS or IDENTIFIERS (Technically meaningful terms or short phrases that characterize a document and could be helpful in cataloguing the document. They should be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location may also be included. If possible keywords should be selected from a published thesaurus, e.g., Thesaurus of Engineering and Scientific Terms (TEST) and that thesaurus identified. If it is not possible to select indexing terms which are Unclassified, the classification of each should be indicated as with the title.)

Unsupervised models of lexical semantics; training corpora; automated corpora generation; context-specific; Latent Semantic Analysis (LSA)