

SIMFAST README file

Michel Génèreux
Natural Language Technology Group
University of Brighton

October 11, 2007

/*

This file is part of SimFast.

SimFast is free software: you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation, either version 3 of the License, or (at your option) any later version.

SimFast is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details.

You should have received a copy of the GNU General Public License along with Foobar. If not, see <<http://www.gnu.org/licenses/>>.

*/

This software computes the semantic similarity (0 is dissimilar, 1 is very similar) between two sentences. It can also be used as a POS tagger. The algorithm implemented for the POS tagger is the Brill tagger, where lexical and contextual rules are applied to a string. To compute similarity, a modified version of the Levenhstein edit-distance algorithm is used where words, not letter, are token. Three edit operation have diffreent cost: insert (cost 1), delete (cost 1) and substitute (max cost is 2). When we substitute two synonyms, the cost is 0 while substituting two antonyms or

when one word is not in the dictionary (WordNet) the cost is 2. When two words are in the dictionary, we compute their similarity using their glosses and the Lesk method. The overall similarity is normalized to the maximum cost. If used to compute similarity, it requires WordNet 2.1 installed. See `wn.h` for paths.

For cultural heritage the tool can be used in question-answering to select the most likely query from a user, see option `-c` below. More information can be found in the file `qa_protocol.doc` with the distribution.

Usage:

To compute similarity of two strings:

```
SimFast -s[v] string1 string2
```

```
e.g. SimFast -s "The cat chases the dog." "The dog chases the cat."
```

```
Similarity is: 0.37931
```

To compute similarity of two files:

```
SimFast -sf[v] file1 file2
```

```
e.g. SimFast -sfv apples.txt oranges.txt
```

To tag the words of string with their POSs:

```
SimFast -t string
```

```
e.g. SimFast "The cat chases the dog"
```

```
The/DT cat/NN chases/VBZ the/DT dog/NN
```

To tag the words of a file with their POSs:

```
SimFast -tf file
```

To compute the most similar strings from a file of questions:

```
SimFast -c question_file answer_file question
```

e.g. SimFast -c questions.txt answers.txt "Was the palace fortified?"

To get help

SimFast -h

Copyright

WordNet Release 3.0

This software and database is being provided to you, the LICENSEE, by Princeton University under the following license. By obtaining, using and/or copying this software and database, you agree that you have read, understood, and will comply with these terms and conditions.:

Permission to use, copy, modify and distribute this software and database and its documentation for any purpose and without fee or royalty is hereby granted, provided that you agree to comply with the following copyright notice and statements, including the disclaimer, and that the same appear on ALL copies of the software, database and documentation, including modifications that you make for internal use or for distribution.

WordNet 3.0 Copyright 2006 by Princeton University. All rights reserved.

THIS SOFTWARE AND DATABASE IS PROVIDED "AS IS" AND PRINCETON UNIVERSITY MAKES NO REPRESENTATIONS OR WARRANTIES, EXPRESS OR IMPLIED. BY WAY OF EXAMPLE, BUT NOT LIMITATION, PRINCETON UNIVERSITY MAKES NO REPRESENTATIONS OR WARRANTIES OF MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE OR THAT THE USE OF THE LICENSED SOFTWARE, DATABASE OR DOCUMENTATION WILL NOT INFRINGE ANY THIRD PARTY PATENTS, COPYRIGHTS, TRADEMARKS OR OTHER RIGHTS.

The name of Princeton University or Princeton may not be used in advertising or publicity pertaining to distribution of the software and/or database. Title to copyright in this software, database and

any associated documentation shall at all times remain with Princeton University and LICENSEE agrees to preserve same.