Automatic Concept Map Generation from Text-Based Learning Material

Documentation

Submitted By

Ashish Malgawa 17CS60R81
Hussain Jagirdar 17CS60R83
Abhishek Prakash Singh 17CS60R48
Shah Smit Ketankumar 17CS60R72

Guided by

Prof. Plaban Kumar Bhowmick

Master of Technology

In

Computer Science

Department of Centre for Educational Technology
Indian Institute of Technology, Kharagpur
April 2018

Contents

Downloads

1. Glove Dataset

GloVe is an unsupervised learning algorithm for obtaining vector representations for words. Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space.

Pre-trained word vectors.

Download Link : ***http://nlp.stanford.edu/data/glove.6B.zip***

This data is made available under the Public Domain Dedication and License .

(For convenience extract this zip file under data directory)

2. Solr 7.1.0

Solr is the popular, blazing fast open source enterprise search platform from the Apache Lucene project. Its major features include powerful full-text search, hit highlighting, faceted search, dynamic clustering, database integration, rich document (e.g., Word, PDF) handling, and geospatial search. Solr is highly scalable, providing distributed search and index replication, and it powers the search and navigation features of many of the world's largest internet sites.

Download Link : http://archive.apache.org/dist/lucene/solr/7.1.0/

3. Pysolr

Pysolr is a lightweight Python wrapper for Apache Solr. It provides an interface that queries the server and returns results based on the query.

Install: $ pip install pysolr

4. Pyspotlight

Pyspotlight is a thin python wrapper around DBpedia Spotlight .

Install: $ pip install pyspotlight

5. Gephi

Gephi is an open-source network analysis and visualization software package written in Java on the NetBeans platform.

Download Link : https://gephi.org/users/download/

6. Core NLP Package

Stanford CoreNLP provides a set of human language technology tools. It can give the base forms of words, their parts of speech, whether they are names of companies, people, etc., normalize dates, times, and numeric quantities, mark up the structure of sentences in terms of phrases and syntactic dependencies, indicate which noun phrases refer to the same entities, indicate sentiment, extract particular or open-class relations between entity mentions, get the quotes people said, etc.

Download Link *:*

http://nlp.stanford.edu/software/stanford-corenlp-full-2018-02-27.zip

Installation Manual

Solr Indexing

Steps:

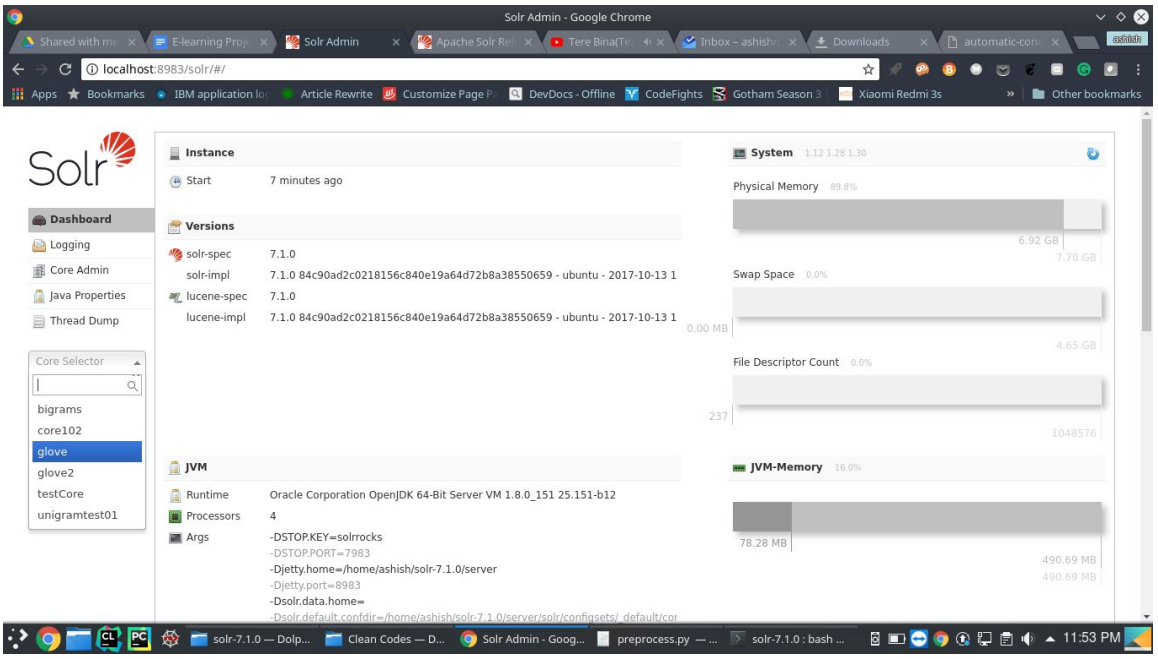
1. Follow this guide to install solr:
   https://lucene.apache.org/solr/guide/7_1/installing-solr.html
2. Open the folder in which you've installed solr.
3. Start solr
   >bin/solr start
   The line below indicates you've successfully installed solr
   Waiting up to 180 seconds to see Solr running on port 8983 [\]
   Started Solr server on port 8983 (pid=24451). Happy searching!

4. Create a core named glove.
   > bin/solr create_core -c glove
   This will come in form of a success message
   WARNING: Using _default configset. Data driven schema functionality is enabled by default, which is
   NOT RECOMMENDED for production use.

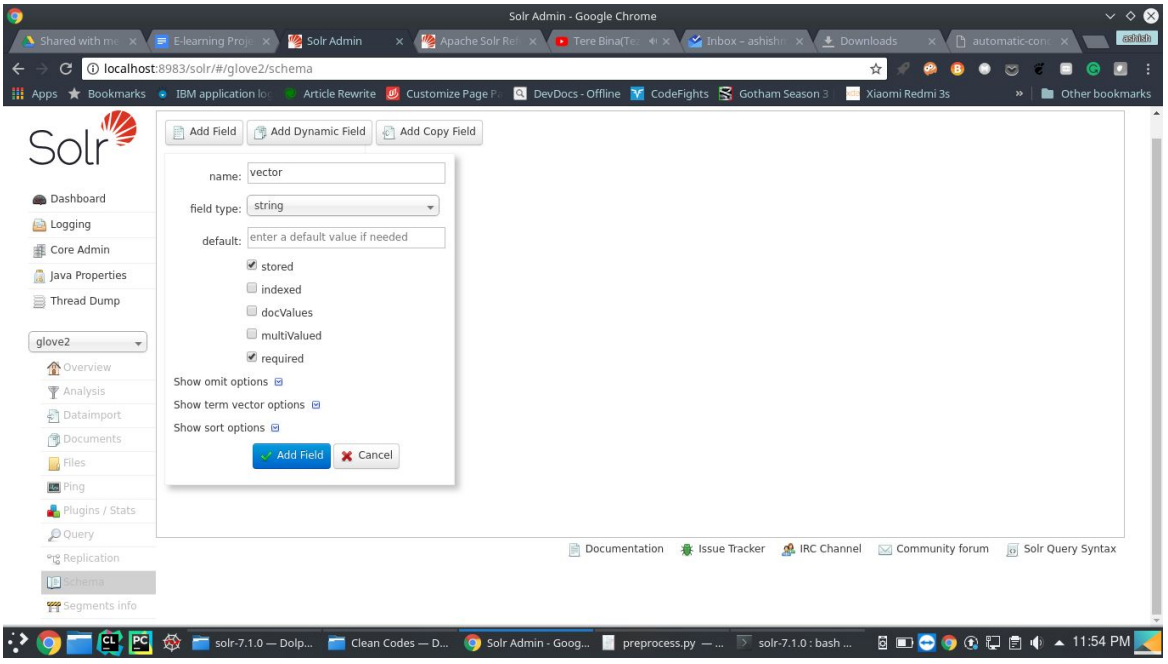
   To turn it off:
   curl http://localhost:8983/solr/glove2/config -d '{"set-user-property": {"update.autoCreateFields":"false"}}'

   Created new core 'glove'
5. Open solr admin panel.

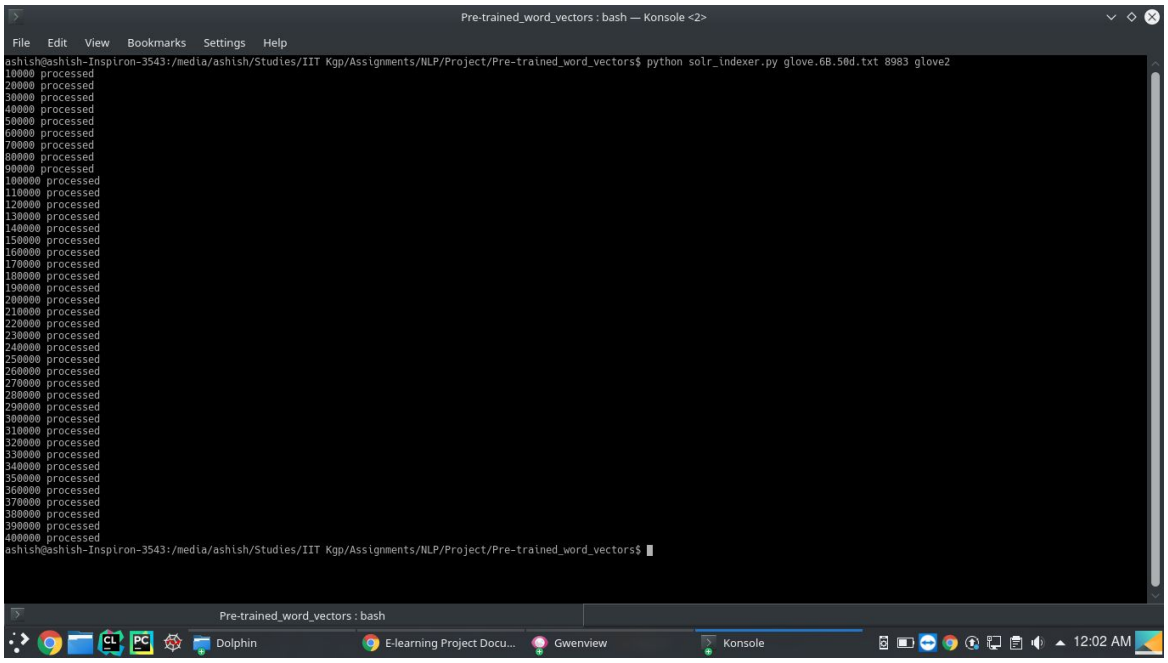
a. Go to this address localhost:PORT_NUMBER_OF_SOLR/solr/



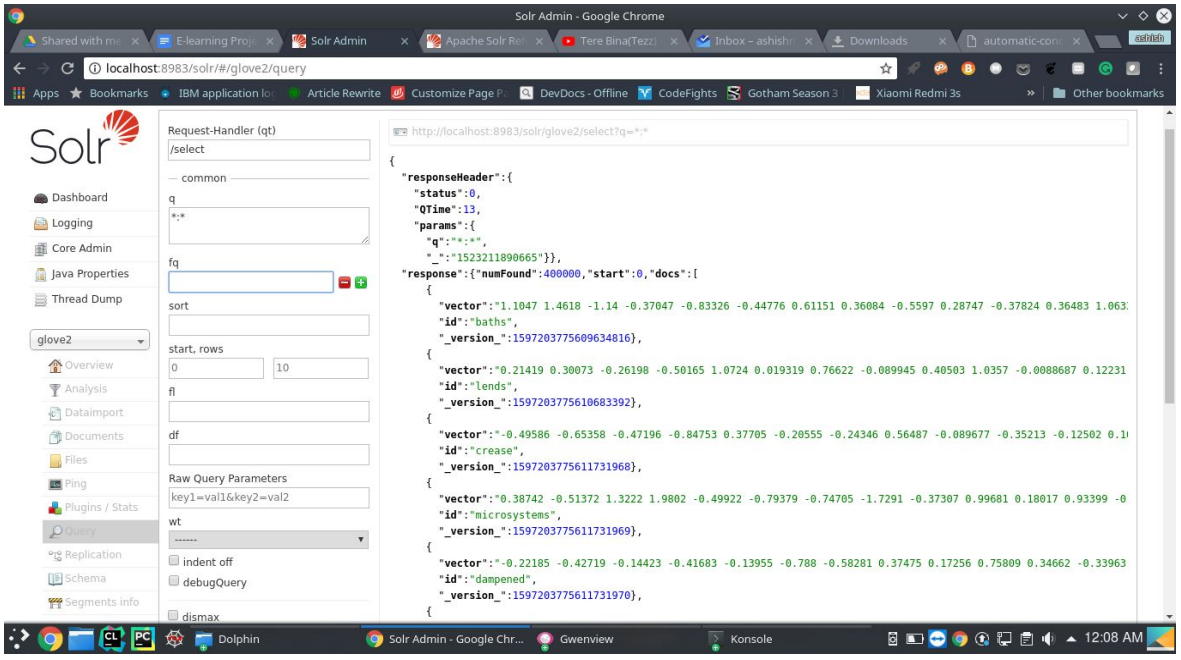
6. Add a string named “vector” field under the schema section.



7. Install pysolr
   > pip install pysolr
8. Run solr_indexer.py with following format
   python solr_indexer.py path_to_glove_dataset PORT_NO CORE_NAME
   Eg. : python solr_indexer.py ../data/glove.6B.50d.txt 8983 glove2
   On success:



To further check the indexed data set go solr admin and press enter on fq



9. Indexing Complete. Happy Searching!

Coreference Resolution and OpenIE

Steps -

1. Download Stanford CoreNLP package from here.
   Or use the following commands:
   wget http://nlp.stanford.edu/software/stanford-corenlp-full-2018-02-27.zip
   Or get using curl:
   curl -O http://nlp.stanford.edu/software/stanford-corenlp-full-2018-02-27.zip

2. Unzip the release:
   unzip stanford-corenlp-full-2018-02-27.zip
   Store the unzipped file in **data** directory of the project.

3. Enter the newly unzipped directory:
   cd stanford-corenlp-full-2018-02-27

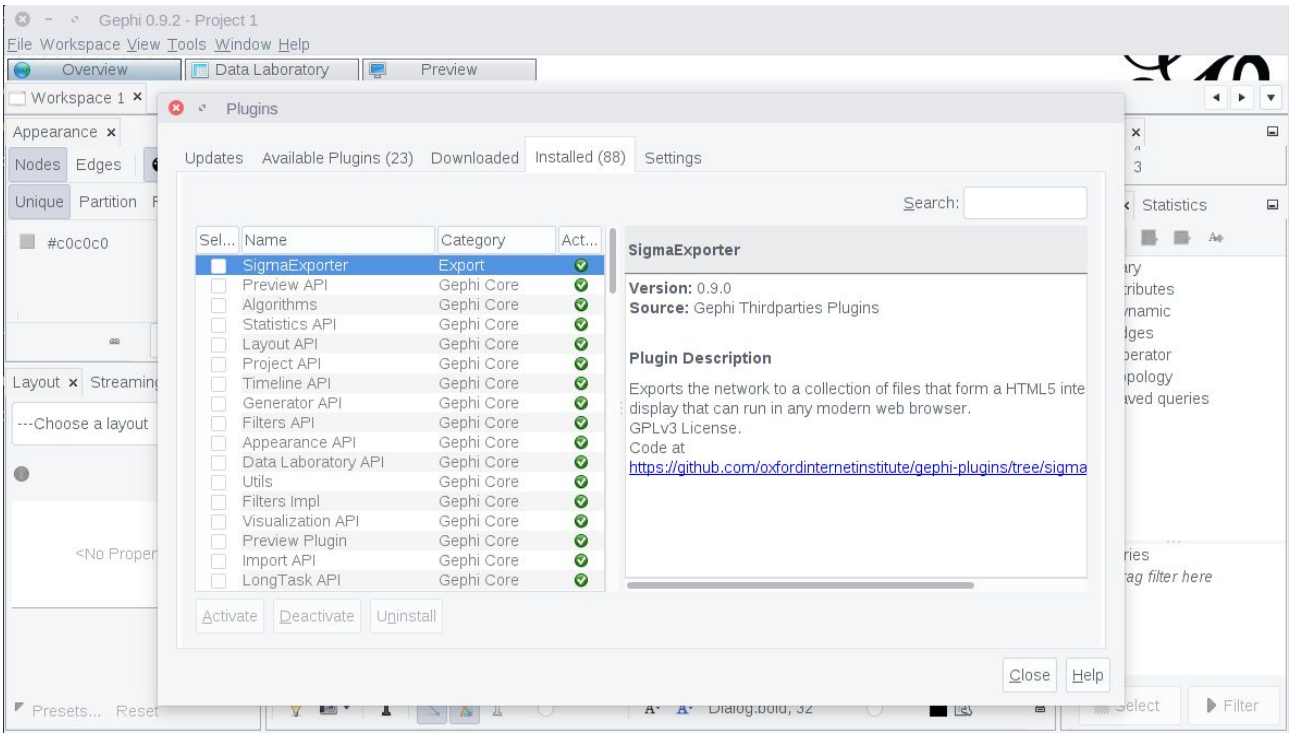
4. Set up your classpath. If you're using an IDE, you should set the classpath in your IDE. If you are using bash or a bash-like shell, the following will work.

   for file in `find . -name "*.jar"`; do export CLASSPATH="$CLASSPATH:`realpath $file`"; done

Gephi Installation

Steps:

1. Follow this guide to install Gephi https://gephi.org/users/quick-start/
2. Unzip the folder.
3. Go to the bin folder and run ‘./gephi’ command to launch the gephi tool. (for ubuntu)



4. Go to tools menu and select Plugins.
5. Download Graph Streamer Plugin.

User Manual

- All code should be present in src folder, data in data folder and output in output folder respectively.
- Run makefile.sh with version number (1.0, 2.0, 2.1, 2.2) as first argument and filename (path : data/<filename>) as second argument.
- Now output would be a gephi file, stored in the output folder.
- Open that gephi file in Gephi tool.
- To display the labels, go to the data laboratory panel (edges) and copy 'hase' field to 'Label' field.
- For other stuff like animation and changing the UI, you can refer the quick guide link. https://gephi.org/users/quick-start/
- Algorithms are explained in the presentation with results.

Evaluation metrics

1. We considered time taken to create a concept map considering number of words in the document. (To check the scalability of the model)

2. **Precision**: We can have a standard hand drawn concept map, and we can compare how close our concept map is from the standard one.

3. **Concept Density**: We can also calculate the concept density for different documents with same length.