Web Mining

Assignment 5: Link News Entities with Wikipedia (5% weightage)

Date Posted: Oct 21, 2013
Date of submission: Nov 13, 2013. 9pm.

Goal: To make students understand the task of entity linking and also to compare richness of data in news versus Wikipedia.

Task: We will provide you with half a month (June 1-15, 2012) worth of news feeds crawled from various news websites. We will also provide you with the Wikipedia dump. Your aim is to link entities in news feeds with Wikipedia entity pages.
The steps for this task are as follows:
1. Find news title from news feeds. The feed pages contain a lot of other details, ignore them. Just get the news titles.
2. Run NER and find entities from news titles.
3. Given the Wikipedia dump, gather all the entities from Wikipedia.
4. Link the entities discovered in steps (2) and (3). Perform normalization where required (like lower casing, inserting/removing spaces or hyphens or brackets etc). If an entity is marked as ambiguous on Wikipedia, resolve the ambiguity by using the context in news title.
5. Find percentage of new entities in news which are not present in Wikipedia.
6. Also extract the timestamp, location, author, website URL of the news (if available) from the news feed.

Dataset:
Wikipedia Data set is available at DC Nick : XYZ@Lab.
Folder : Wiki
filename : enwiki-latest-pages-articles.xml (42 GB) also contain a sample file (100 MB)

News data is available on the course portal assignments page as an attachment.

Submission Instructions: Create a directory with the name "<rollno>_as5".
Within that you need to put in the following files: The file entityLinkage.txt, a readme file README.txt and your code directory zipped as code.zip.

entityLinkage.txt should have the following format per line. On each line, you need to have 8 values separated by a tab.

1. Filename from which the news is extracted
2. News title
3. Entities discovered from news title separated by spaces
4. Wikipedia URLs for those entities separated by spaces.

5. Timestamp for news (if available)
6. Location of news (if available)
7. Author or news agency of news (if available)
8. Website URL of the news (if available)

The README.txt should

1. Describe your methodology in brief.
2. What heuristics you used to disambiguate the entities
3. What steps were taken to normalize the news entity set and the Wikipedia entity set for good matching
4. Which NER package you used
5. How many entities present in news are not present on Wikipedia
6. Instructions about how to compile and run your code.

The code directory should contain all your code. Zip the code directory to create code.zip

Finally, zip the "<rollno>_as5" directory to get <rollno>_as5.zip and submit the zip file.