IN4252 Web Science & Engineering
Hands-on Assignment - 2

Ke Tao (k.tao@tudelft.nl)

December 3, 2013

1 Dataset

1.1 Get the dataset

For each of the students that have already enrolled on Blackboard, I have created a dataset. You can get your own one by accessing following link:
`http://www.st.ewi.tudelft.nl/~ktao/wse2013/handson-2-data.html`

1.2 Description of the dataset

In the dataset, you will find tweets posted by 10 users during a period of two months. Please **DO NOT** spread the dataset for any purposes other than this assignment. All of the users in the dataset contributed at least 20 tweets in total and at least one tweet in each month of our observation period.

1.3 Import into Database

Given you have done the preparing assignment, I think there is no reason to be explain the approach again.

Please first create a database, and then import the sql dump file that contains the schema and data in a table named tweets_sample.

2 Task 1: Semantic Extraction

You may use Named Entity Recognition Services to process a textual snippet, and you can get a list of identified Named Entities. There are four such services that we know. You are recommended to try all of them. I personally recommend to use OpenCalais for the reason of stability.

2.1 OpenCalais

You need to register for getting an API key (`http://www.opencalais.com/APIkey`) to use this service. The documentation can be found at :

`http://www.opencalais.com/calaisAPI`

For example, if you want to use OpenCalais as a REST service by encoding parameters in the URL, you may want to have a look at this:

`http://www.opencalais.com/documentation/calais-web-service-api/api-invocation/rest-using-paramsxml`

Specifically, if you use Java, Python, or PHP, you can make use of codes from following projects:

Java `https://code.google.com/p/j-calais/` → pay attention to the dependencies.

Python `https://code.google.com/p/python-calais/`

PHP `https://github.com/dangrossman/PHP-OpenCalais`

2.2 DBpedia Spotlight

This is a product mainly made by researchers from Free Berlin University. You can find the open-source project via this link:

`https://github.com/dbpedia-spotlight/dbpedia-spotlight`

However, we are not going to compile it and set up a server on our own machine because it costs a lot of hardware resources. You may want to try this service on the demo site:

`http://spotlight.dbpedia.org/demo/`

More detailed information can be found at :

`https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki`

2.3 Zemanta

Zemanta was designed to be an assisting tool for composing blog entries. It can identify the named entities when you are writing blog posts, and automatically annotates the entities with the hyperlinks to the web pages that can describe them in detail. What's more, it can recommend you to add related pictures, related articles to your blog posts. You can find a demo at:

`http://www.zemanta.com/demo/`

You also need an API key to use this service. The documentation can be found at:

`http://developer.zemanta.com/docs/`

The sample code are also provided by Zemanta. You can find them at:

`http://developer.zemanta.com/wiki/`

2.4 AlchemyAPI

AlchemyAPI is an option not only for Named Entity Recognition. It also support sentiment analysis, language detection etc. You also need to get an API key to use this service. Free API key grants you 1,000 call per day. Therefore

it might be a business solution when you want to pay in order to get a stable NER service. You can try the demo at:
`http://www.alchemyapi.com/api/demo.html`
The documentation can be found at:
`http://www.alchemyapi.com/api/entity/`

3 Task 2: User Modeling and Recommendation

In this task, your will design and implement functionality that provides a personalized recommendation that recommends tweets to a user. Therefore, you should engineer (a) user modeling functionality that allows for inferring the interests of a user and constructing the user profiles and (b) recommendation functionality that - given a user profile and a set of candidate tweets - allows for ranking the items so that the items which are most relevant to the user (profile) appear at the top of the ranking.

The key challenge of this assignment is to design the user modeling functionality that allows for computing personalized tweet recommendations. Therefore the following steps need to be taken:

1. Design the user modeling functionality that infers the interests of a user and further constructs the user profile.
2. Design the recommendation functionality: in the lecture, we introduced two basic classes of recommendation approaches - content-based and collaborative filtering-based. You need to choose one of them, which is suitable for your application.
3. Implement your user modeling and recommendation functionality.
4. Apply your user modeling and recommendation functionality on the given Twitter data to produce user profiles and recommendations.
5. **[Optional]** Evaluate the quality of the recommendation using at least one metric. You can find more information regarding how to evaluate a recommender system in some research articles listed in the Section Pointer.
6. Report on your application.

4 To be delivered

The deliverable of this assignment should be in pdf format and named as "[Last name].[First name].pdf". You are expected to deliver the following artifacts for this assignment:

1. **Report**. A 1-2 pages summary of your application that

(a) describes how the user modeling work, i.e. given the Twitter activities (data), how your user modeling functionality infers the interests of a user and construct user profiles.

(b) describes which recommender approach you have selected and why.

(c) describes the specifics of the implementation.

(d) [**Optional**] discusses strengths and weaknesses/limitations of your approach and "future work" - how you can further improve your current approach?

(e) [**Optional**] describes the metrics you have chosen to measure the quality of the recommendations and the evaluation results.

2. **Example user interest profiles and recommendations**. For 3 sample users, provide a short description of user interest profiles (e.g. list of (top) topics they are interested in) and the top 3 tweets that are actually recommended to the users. The examples should be into the appendix of your report.

   [**Optional**] If you want to discuss strength/weakness of your application, it could be wise to carefully select/specify the example user profiles (e.g. the examples can be used to illustrates strengths and weaknesses of the proposed user modeling and recommendation functionality).

5 Pointers

5.1 Suggested research articles

- Abel, F., Gao, Q., Houben, G.-J., Tao, K.: Analyzing User Modeling on Twitter for Personalized News Recommendations. In Proceedings of the International Conference on User Modeling, Adaptation and Personalization (UMAP), Girona, Spain, Springer (2011)

  `http://www.st.ewi.tudelft.nl/~ktao/wse2013/twiterum-umap2011.pdf`

- Chen, J., Nairn, R., Nelson, L., Bernstein, M., Chi, E.: Short and tweet: experiments on recommending content from information streams. In: Proceedings of the International Conference on Human factors in Computing Systems (CHI), New York, NY, USA, ACM (2010)

  `http://www.parc.com/publication/2400/short-and-tweet.html`

5.2 Other resources

1. Mahout[1]: a library that can help you implement some similarity metrics and recommendation algorithm.

[1] http://mahout.apache.org/

2. RecSysWiki[2]: a wiki for sharing information related to recommender system.

6 Deadline

Please submit your homework before **December 16th, 2013, 23:59 (CET)**.

7 Q&A

If you have any question, make sure you have read this document and the FAQ web page via following link:

`http://www.st.ewi.tudelft.nl/~ktao/wse2013/faq-handson-2.html`

If you still don't get the answer after that, please send an email and good luck!

[2] http://recsyswiki.com