Initial Plan

"Identifying narrative text on reddit"

Cardiff University School of Computer Science and Informatics
CM3203 – One semester individual project – 40 Credits
Student: Adriano Sole
Supervisor: Steven Schockaert

Contents

Project Description

Neural network models for natural language processing are typically pre-trained on large text collections. This allows these models to learn world knowledge without the need for an explicit supervision signal. However, the kind of knowledge they can learn in this way crucially depends on the type of text collection that is used. For instance, Wikipedia is a common choice. By pre-training models on encyclopaedic text, they can acquire a lot of factual knowledge about the world. On the other hand, training models on narrative text (e.g. books or movie scripts) can be a better choice if learning common sense knowledge is the main goal.

In recent years, the social media site Reddit is increasingly being used for training natural language processing models. However Reddit contains documents covering a broad range of genres, including expository, narrative, and argumentative text.

The aim of this project is to develop a method for identifying narrative documents in a large collection of Reddit posts, roughly 1000 to begin, with more posts being needed towards the second half of the project.

During this project I will need to learn how to utilise multiple different natural language processing tools and models, particularly the models mentioned in the paper “Identifying Personal Stories in Millions of Weblog Entries” (Gordon & Reid, 2009) as well as the word vector model in “Efficient Estimation of Word Representation in Vector Space” (Mikolov, Chen, Corrado, & Dean, 2013).

The model will be built using the Python coding language, as it is simple and contains many NLP tools that I have experience with in previous projects, such as the Natural Language Toolkit (NLTK) developed at the university of Pennsylvania (Bird, Klein, & Loper, 2019).

Further to this I will also be required to gather the data needed to train the natural language processing model that I will develop. In order to gather the Reddit data I will be learning how to use the Reddit API in order to store the data.

Aims and Objectives

The following aims and objectives will be separated into two different sections, one section for the first goal of the project and one for the second goal.

First goal

The initial goal of the project to be enacted during weeks 3-5 will be to develop a working model of what is proposed in the paper “Identifying Personal Stories in Millions of Weblog Entries” (Gordon & Reid, 2009). In said paper Gordon and Reid found that there were many prominent features that narrative weblog posts contained which differed from the non-narrative posts. I will be following the method they used by manually classifying a collection of Reddit posts as narrative or non-narrative to be used as training data for the system.

The goals of this section can be summarised as follows:

- Develop a method for correctly identifying narrative Reddit posts using the above paper as a guide
- Gather statistics of the quality of the identifications using a sample size of roughly 1000 reddit posts from the Cardiff subreddit (https://www.reddit.com/r/Cardiff/) as well as posts from a range of other subreddits

- Try to achieve a working system that can successfully classify a post as a narrative 25% of the time
- Begin the data collection of Reddit posts that will be used as training data for the system, having roughly 1000 posts collected by the start of week 4

Second goal

After the completion of the first goal I should have a working method for identifying narrative documents based on the above mentioned paper. The following goal I have set for the 3 weeks following the completion of the first goal, weeks 6-8, wherein I will be able to attempt to further improve on the system that I will have created by involving the use of word vectors with the Word2vec technique originally created by a team of researchers at Google (Mikolov, Chen, Corrado, & Dean, 2013). In addition, I will be using data gathered from multiple different subreddits in order to provide training data for the model. Using this extra data I will be able to test the model I have created to automatically detect if a subreddit is likely to contain narrative posts, which will allow me to train a classifier based on the subreddits that are likely to contain narrative text.

The goals of this section can be summarised as follows:

- Develop the first goal further, upgrading the system to utilise word vectors
- Show an improvement in the accuracy of the system being able to classify texts as narratives
- Acquire extensive comparative data from the system in its final state, to be compared with the data in the initial goal
- Train a classifier using data gathered from a number of different subreddits
- Try to achieve an improved system that can successfully classify a post as a narrative 50% of the time

Ethics considerations

Due to the primary focus of this project requiring the analysis of Reddit posts, there is an ethical issue to the work that will be required once development starts. It will be important to obtain ethics approval first before doing any data collection.

As all of the Reddit posts will be taken from a number of public subreddits there will not be a huge issue with the scraping of this data from the web. However it will still be important to consider the possible ethical issues involved with storing the data.

To avoid any issues with data storage I will be storing the data on the Cardiff University OneDrive in order to be compliant with Cardiff University Information Security policy.

When gathering the data I will ensure that only the text from within the post will be kept, which will ensure the anonymity of any users' data involved in the project. This will avoid any issues that could arise from storing personal information of users without their consent as the text that will be stored will not contain any personal information that could be linked back to a specific user.

Work plan

Week 1

Initial plan and background research into natural language processing.

This will provide me with a general background as to what the project will be about and give some initial reading into the methods that I will be using when I begin working on the development in week 3.

Week 2

Further background research into natural language processing, with a focus on the different types of NLP tools and methods of analysing language.

At this point I will be ready to begin gathering data and developing the code for the system. I will understand enough about the NLP system mentioned in the "Identifying Personal Stories in Millions of Weblog Entries" paper (Gordon & Reid, 2009) and also have a general idea of how to implement word vectors when it comes to week 6.

Weeks 3 – 5

During these 3 weeks I will be working on the first goal mentioned above, to implement a working system that can classify a Reddit post as a narrative. Additionally during the early weeks I will also be gathering data to be used in the system itself by using a Reddit post scraping tool.

By the end of week 5 I will have a working system that can classify a Reddit post as a narrative at least 25% of the time.

Weeks 6 - 8

These next 3 weeks will be devoted to improving on the initial goal, I will be attempting to implement the Word2Vec model into the developed system to see if there is an improvement in the accuracy of the natural language processing model. Additionally I may be able to expand the training data to more than just one subreddit by gathering more posts from different areas of Reddit.

By the end of these 3 weeks I will have a developed system that can classify a Reddit post with an accuracy of roughly 50%.

Easter weeks 1-3

Tentatively free, some room for additional overhead from previous goals.

Week 9

Evaluation of results towards first and second goals during weeks 3-8.

Week 10

Final report: Summarise initial goal and report results for weeks 3-5

Week 11

Final report: Notes and results from second goal in weeks 6-8

Week 12 – 10/05/21

Finalise report towards the beginning of week 12 and use the rest of the week and submit for deadline on 14th of May

Works Cited

Bird, S., Klein, E., & Loper, E. (2019). *Natural Language Processing with Python.* Pennsylvania: O'Reilly Media. Retrieved from NLTK documentation website: https://www.nltk.org/book/ch00.html

Gordon, A. S., & Reid, S. (2009). *Identifying Personal Stories in Millions of Weblog Entries.* California: University of Southern California.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space.* California: Google.