Final Year Project Interim Report

A Word Lexicon Builder Using Neural Word Embeddings

Daire Murphy - 15441458

Supervisor: Derek Greene

A thesis submitted in part fulfilment of the degree of
BSc. (Hons.) in Computer Science with Data Science

UCD School of Computer Science
University College Dublin

Table of Contents

List of Figures

List of Important Abbreviations Used Within

TLA	*Three Letter Acronym*
FLW	*Four Letter Word*
PFO	*Public Funding Option*

Abstract

Word lexicons are commonly used to systematically filter and search large text corpora, to isolate all documents that are related to the selected concept of interest. Using a combination of pretrained word embeddings, as well as self-developed word recommendation methods, this project evaluates and compares the word recommendations produced from each of these models. This data is examined to identify the most accurate models to be used. An interactive web interface is implemented, using the purposed word embeddings to create a lexicon recommendation method. This web interface allows for editing of recommended words, improving accuracy of the models for users. Visualizations of the embedding spaces are added to the interface to allow exploration by users of the word embedding models used within this project.

Project Specification

Word embeddings, which refer to a set of language modeling and feature learning techniques in natural language processing algorithms that map words or phrases from a vocabulary to vectors of real numbers, have over the last number of years been used in the creation of word lexicons, a list of words interested to a specific theme or concept, on large text corpora. Word embedding models are trained by applying a neural network to a large text dataset, and such many pretrained models are available on an extreme variety of datasets.

A primary focus of this project was the creation, development and evaluation of several word recommendation methods, a long with further comparison and evaluation of these models with other pretrained models. The implementation of these models to an interactive web interface is required to create an online lexicon recommendation method. The student will then be expected to add interactive visualizations to the web interface to allow exploration of the embedding spaces by users

Core:

- Develop a word recommendation method, based on a given seed set and a single word embedding model.
- Perform an evaluation to compare the word recommendations produced using embedding models trained on different text datasets and using different algorithms.
- Implement an interactive web interface for lexicon building, which uses the proposed lexicon recommendation method. This interface should allow users to accept or reject recommended words, in order to produce a more useful final lexicon.
- Develop an “ensemble” word recommendation method, which combines the outputs from different embeddings into a single set of recommendations. Incorporate this new method into the web interface.

Advanced:

- Add interactive visualisation functionality to the web interface to allow users to explore the embedding spaces in more detail.
- Design and conduct a user study to assess the usefulness of the web interface and the lexicons which it produces.
- Perform an evaluation to examine the extent to which embedding algorithms can produce different results when applied to the same data.

Chapter 1: **Introduction**

Introduce your vision of the project here. Describe the domain of the project, and the intended application. A well-written report will answer three key questions: *What* am I doing in this project? *Why* is it worth doing? *How* do I plan to go about it? In this introductory section, offer a concise answer to the *What*, and follow-up with a compelling account of the *Why*. Leave the *How* to a subsequent section. Do not try to do too much in any single section of the report. By providing details in a logical order, you will show that you have a plan for the report *and* the project.

1.1 Motivations

1.2 Project Objectives

1.3 Summary of Report

Chapter 2: **Related Work**

A key task of this first report is to establish a baseline against which your later work will be judged. Your FYP project does not exist in a vacuum, and its central problem, or a variant thereof, will have been tackled by others before you. In this section, you should describe how previous approaches have tackled the problem, and clearly articulate the state of the art (or SOA) for your project.

For research-oriented projects, this task will be time-consuming but relatively straightforward. You should read past works on the subject, summarize the main points, pros *and* cons, and root out the previous works that they cite in turn. You may use Wikipedia as a *secondary* source only, which is to say that it can be a useful first port of call on many topics but not a source that should be liberally cited. Rather, use Wikipedia as a hub for gathering references to primary work in the field (original papers and reports), then read and summarize those. Do not quote a work that you have not read, unless you are quoting someone else's view of that work. Never use another writer's words as your own. Place any extracts from another's work in double quotes, and attribute the quotation to its author with a citation. It is a very low act to plagiarize another's work and take credit for their words, so tread carefully. Even unintentional plagiarism is still plagiarism.

For more application-oriented projects, you are still expected to survey other solutions, either for the given problem or for similar problems, and also consider applications that share functionality or design principles with your own. In short, this section is the core of your report regardless of what kind of project you do.

2.1 Text Analysis

2.2 Word Embeddings

2.3 Lexicon Building

Chapter 3: **Progress to Date**

3.1 Data Considerations

In this section you should characterize the nature and scale of the data you are working with. Outline the shape of the data, where you expect to obtain it, and the size of the data. Is it static or dynamic, local or remote, stored or streaming? Is it raw or structured? Is it unfiltered user data, or is it curated by a specialist? What is your rationale for using *this* data and not other data? If your project looks at callout times for Spanish ambulances, usage rates of French parking lots, alcohol consumption in Germany, and so on, then explain why you are not using Irish data for the project. Indicate the data-cleaning processes that you anticipate will be necessary. What licensing restrictions, if any, apply to your data? Will you be making this data public after your project is completed? Are there any privacy or ethical issues with how the data is to be collected or used? If so, discuss here.

Some or many of these questions may be moot in the case of specific projects, but you should provide compelling answers to any that seem relevant. Since this provides the foundation for your project, your reviewers will be looking closely.

3.2 An Outline of My Approach

In this section present an outline of your considered approach to the problem at the centre of your project. Clearly present your design choices, or your choice of algorithms, and any pertinent model parameters. For instance, if you plan to use a genetic algorithm, outline here a sense of your fitness function, major variables, population size, and so on, so that your reviewers can critique your choices. If you opt for a neural architecture, describe your chosen framework, and motivate the number and kinds of layers in your network. In short, be specific about the choices you are committing to at this stage. Being vague and non-specific will not help your case, as your report will be graded in large part on the specificity and perceived wisdom of your choices. Remember also that feedback is intended to help you as you progress to the next stage of your project. If you give reviewers little to chew on, they will not be able to give you specific feedback and guidance.

Chapter 4: **Future Work**

4.1 Project Work Plan

In this section you will present a work plan for the remainder of your project. Show that you have considered the issues carefully, and that you can be trusted to lead a research or development effort. Be as specific as you can about the time you expect to allocate to each work component, and the dependencies they have to each other. A Gantt chart is helpful in this respect but do show some sense in how you present your plan. A naïve understanding makes for a simplistic plan.

A key part of a successful project is *evaluation*. It is not enough to just state that your project is a success, or that your friends seem to like it. You must have a plan for evaluating the end result. How you evaluate will depend on the nature of your project, and you should have a serious conversation with your supervisor about evaluation before you get to this stage. Will your work yield quantitative results that can be compared to past work or to established benchmarks? Does your work consider different configurations of a system or a solution that you can compare to each other, allowing you to empirically find the best one? Do you have a sample user pool for your planned application, and are they willing to give you structured qualitative and quantitative feedback (e.g. via a questionnaire)? However, you plan to evaluate your project, please sketch your intentions here.

Chapter 5: **Summary and conclusions**

In this section you will sum up your report, draw some conclusions about your work so far, and make some general observations about the work to come. You may also use this opportunity to express points of view, or make factual claims, that are more pertinent here than in other sections of the report. If your project raises some ethical concerns, for example about how data or users are treated, then address them here in a thoughtful manner.

Regarding this document, here are some concluding points that you should keep in mind when writing your own. You may use screenshots in your report, but do not overfill your report with them, or with figures of any kind. Make sure that figures earn their keep, and are not just present as space fillers or as eye candy. If you use diagrams or figures from other people's work, including the web, be sure to cite the creator in the corresponding caption. All things being equal, it is better to construct your own figures than to copy and paste those of others. In any case, always make sure that your images are readable, do not suffer from pixilation or aliasing effects, and that each is clearly numbered, captioned and meaningfully referenced in the main body of the text.

Ensure that there is a cohesive argument expressed in the text of the report and that it is not simply a bag of diagrams, screenshots and wishful thinking. Every report should tell a story, so know what story you want to tell. When you include images, make sure they are readable and truly add to the discussion.

Make sure your language is professional throughout, and steer a course between pompous and colloquial. Maintain authorial distance and do not overuse "me," "I" and "our." Your are writing for a professional audience who will judge you on the quality of your prose, so use a grammar and a spelling checker.

Use *LaTeX* if you wish – this is recommended if you plan to use mathematical formulae in your report, but in any case, keep the general spacing and font/style you find here (Single or 1.5 spacing, 12 pt. font for text, etc.). Be sure to submit a PDF (never a .DOC or .DOCX file) as your report. If you prepare your report in MS Word, as this document has been, save it as a PDF before you submit it. Overall it should be about **18 – 20 pages**, including figures, front matter and references, A significant portion of the report will be textual, with approx.. five or six thousand words. Do not rely on images or other filler to write your report for you.

The dates and means of submission will be communicated to you separately.

Acknowledgements

Name check any person who helped you with this work. Acknowledge that the work is entirely your own, and that every sentence in this report was written by you and you alone. Plagiarism is a very serious infraction that will be dealt with severely. Avoid any ambiguity on this point by citing things carefully!

References

List any bibliographical citations here for people and work that you quote/cite in the main body of your report. Use the general format below for all bibliographic entries. Ensure each entry is complete (including author, year, title, publication).

Be sparing in your citation of URLs and Wikipedia pages. Do not cite bare URLs unless absolutely necessary – cite instead the print publication if possible.

Dantes, E. (1762). Escaping from tight corners. *Monte Cristo press*, Paris, France.
Squarepants, S.B. and Tentacles, S. (2003). Hygiene Issues In The Crabby Patty. *Journal of Aquatic Foodstuffs*, vol. 7, no. 6, pp 23-32.
Drumpf, D. J. (2016). The Effects of Magnetism on Cats. *Phys. Rev. Letters D*., vol. 203, no. 8, pp 56-59.
Turing, A. M. (2019). The Turing Test: A History of a Misunderstood Idea. *Journal of Paranormal Communications*., vol. 17, no. 18, pp 13-29.