Project 1

Jesus Soto Gonzalez
jhsoto@iastate.edu

Abstract

This first project for COM S 5790, Natural Language Processing applies basic NLP and word embedding methods to a given dataset. We preprocessed the text by tokenizing it, removing stopwords, and using phrase mining. From this, we created frequency- and TF–IDF-based word clouds to highlight key terms. For semantic analysis, we trained Word2Vec models on both words and phrases to find the nearest neighbors of top TF–IDF terms. Phrase mining with `gensim`'s bigram and trigram models helped capture multiword expressions and produced clearer, more meaningful embeddings.

1 Methods

1.1 Preprocessing

The dataset was first filtered to only include abstracts with Category = 1. Each abstract was sentence-split, tokenized, lowercased, and cleaned by removing stopwords and non-alphabetic tokens using NLTK. The result was saved both as plain text for word clouds and as token lists in JSON for Word2Vec tasks respectively. This step ensured consistency across all later tasks and reduced noise from irrelevant or low value tokens.

1.2 Phrase Mining

To capture multiword terms, I used the `gensim` library for phrase mining. I first trained a bigram model on the tokenized corpus with a minimum count of 8 and a threshold of 8.0. This combined pairs of words that often appeared together. Then, I trained a trigram model on top of the bigrams to join longer expressions. In both cases, phrases were written with underscores, such as *carcass_composition*, *average_daily_gain*, and *backfat_thickness*.

I also tested different parameter values. Stricter settings 10, 10.0 created fewer but very strong phrases, while looser ones 5, 5.0 or 3, 3.0 created more phrases but added noise. The balanced setting 8, 8.0 worked best, since it kept important domain phrases like *quantitative_trait_locus* and *meat_quality* without introducing too many random ones. This choice was important because phrase quality directly influenced the TF–IDF terms and the results from Word2Vec.

1.3 TF–IDF and Word Clouds

From both the single word and phrased corpus I computed frequency counts and TF–IDF scores. The top-ranked terms were visualized as word clouds, which made it easy to compare terms that are common against those that are more distinctive. This provided a clear way to see how phrase mining and weighting methods highlight different aspects of the corpus.

1.4 Word2Vec

I trained Word2Vec models on both the words and phrased corpus using CBOW with a vector size of 100, a window of 5, and a minimum count of 10 as required. Then, to evaluate the models, I looked at the nearest neighbors of the top TF–IDF terms and compared how the results changed when using phrases instead of single words.

2 Main Results

2.1 Word Clouds – Task 1

The frequency-based word cloud was dominated by very common research terms such as *qtl*, *traits*, *study*, and *genes*. These words appear often across many abstracts but do not point to specific findings. In contrast, the TF–IDF word cloud highlighted more distinctive domain terms, including *ascites*, *earlobe*, *pleurisy*, and *ketosis*. This shows how TF–IDF can surface technical vocabulary that is less frequent overall but more informative for understanding the dataset.

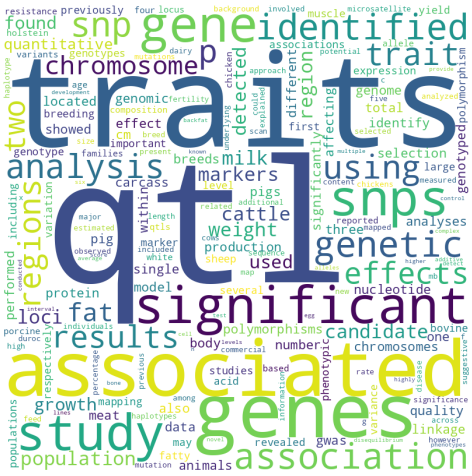

Figure 1: Task 1 frequency-based word cloud. Common words such as *qtl* and *traits* dominate.

Figure 2: Task 1 TF–IDF-based word cloud. Distinctive domain terms like *ascites* and *earlobe* stand out.

3 Preamble

The first line of the file must be

```
\documentclass[11pt]{article}
```

To load the style file in the review version:

```
\usepackage[review]{acl}
```

For the final version, omit the `review` option:

```
\usepackage{acl}
```

To use Times Roman, put the following in the preamble:

```
\usepackage{times}
```

(Alternatives like txfonts or newtx are also acceptable.)

Please see the LaTeX source of this document for comments on other packages that may be useful.

Command	Output	Command	Output
{\"a}	ä	{\c c}	ç
{\^e}	ê	{\u g}	ğ
{\`i}	ì	{\l}	ł
{\.I}	İ	{\~n}	ñ
{\o}	ø	{\H o}	ő
{\'u}	ú	{\v r}	ř
{\aa}	å	{\ss}	ß

Table 1: Example commands for accented characters, to be used in, *e.g.*, BibTeX entries.

Set the title and author using `\title` and `\author`. Within the author list, format multiple authors using `\and` and `\And` and `\AND`; please see the LaTeX source for examples.

By default, the box containing the title and author names is set to the minimum of 5 cm. If you need more space, include the following in the preamble:

```
\setlength\titlebox{<dim>}
```

where `<dim>` is replaced with a length. Do not set this length smaller than 5 cm.

4 Document Body

4.1 Footnotes

Footnotes are inserted with the `\footnote` command.[1]

4.2 Tables and figures

See Table 1 for an example of a table and its caption. **Do not override the default caption sizes.**

As much as possible, fonts in figures should conform to the document fonts. See Figure 3 for an example of a figure and its caption.

Using the `graphicx` package graphics files can be included within figure environment at an appropriate point within the text. The `graphicx` package supports various optional arguments to control the appearance of the figure. You must include it explicitly in the LaTeX preamble (after the `\documentclass` declaration and before `\begin{document}`) using `\usepackage{graphicx}`.

4.3 Hyperlinks

Users of older versions of LaTeX may encounter the following error during compilation:

[1]This is a footnote.

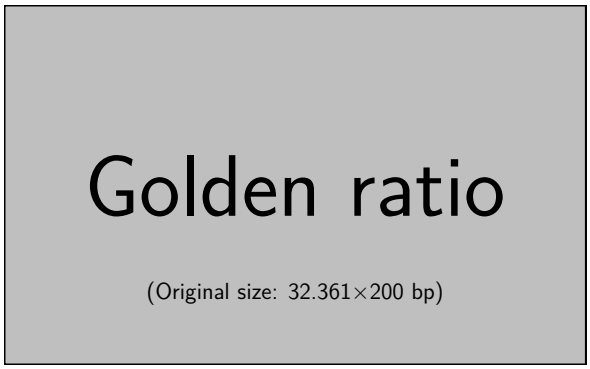

Figure 3: A figure with a caption that runs for more than one line. Example image is usually available through the `mwe` package without even mentioning it in the preamble.

> `\pdfendlink` ended up in different nesting level than `\pdfstartlink`.

This happens when pdfLaTeX is used and a citation splits across a page boundary. The best way to fix this is to upgrade LaTeX to 2018-12-01 or later.

4.4 Citations

Table 2 shows the syntax supported by the style files. We encourage you to use the natbib styles. You can use the command `\citet` (cite in text) to get "author (year)" citations, like this citation to a paper by Gusfield (1997). You can use the command `\citep` (cite in parentheses) to get "(author, year)" citations (Gusfield, 1997). You can use the command `\citealp` (alternative cite without parentheses) to get "author, year" citations, which is useful for using citations within parentheses (e.g. Gusfield, 1997).

A possessive citation can be made with the command `\citeposs`. This is not a standard natbib command, so it is generally not compatible with other style files.

4.5 References

The LaTeX and BibTeX style files provided roughly follow the American Psychological Association format. If your own bib file is named `custom.bib`, then placing the following before any appendices in your LaTeX file will generate the references section for you:

```
\bibliography{custom}
```

You can obtain the complete ACL Anthology as a BibTeX file from `https://aclweb.org/anthology/anthology.bib.gz`. To include both the Anthology and your own .bib file, use the following instead of the above.

```
\bibliography{anthology,custom}
```

Please see Section 5 for information on preparing BibTeX files.

4.6 Equations

An example equation is shown below:

$$A = \pi r^2 \tag{1}$$

Labels for equation numbers, sections, subsections, figures and tables are all defined with the `\label{label}` command and cross references to them are made with the `\ref{label}` command.

This an example cross-reference to Equation 1.

4.7 Appendices

Use `\appendix` before any appendix section to switch the section numbering over to letters. See Appendix A for an example.

5 BibTeX Files

Unicode cannot be used in BibTeX entries, and some ways of typing special characters can disrupt BibTeX's alphabetization. The recommended way of typing special characters is shown in Table 1.

Please ensure that BibTeX records contain DOIs or URLs when possible, and for all the ACL materials that you reference. Use the `doi` field for DOIs and the `url` field for URLs. If a BibTeX entry has a URL or DOI field, the paper title in the references section will appear as a hyperlink to the paper, using the hyperref LaTeX package.

Limitations

This document does not cover the content requirements for ACL or any other specific venue. Check the author instructions for information on maximum page lengths, the required "Limitations" section, and so on.

Acknowledgments

This document has been adapted by Steven Bethard, Ryan Cotterell and Rui Yan from the instructions for earlier ACL and NAACL proceedings, including those for ACL 2019 by Douwe Kiela and Ivan Vulić, NAACL 2019 by Stephanie Lukin and Alla Roskovskaya, ACL 2018 by Shay Cohen, Kevin Gimpel, and Wei Lu, NAACL 2018 by Margaret Mitchell and Stephanie Lukin, BibTeX suggestions for (NA)ACL 2017/2018 from Jason Eisner, ACL 2017 by Dan Gildea and Min-Yen Kan, NAACL

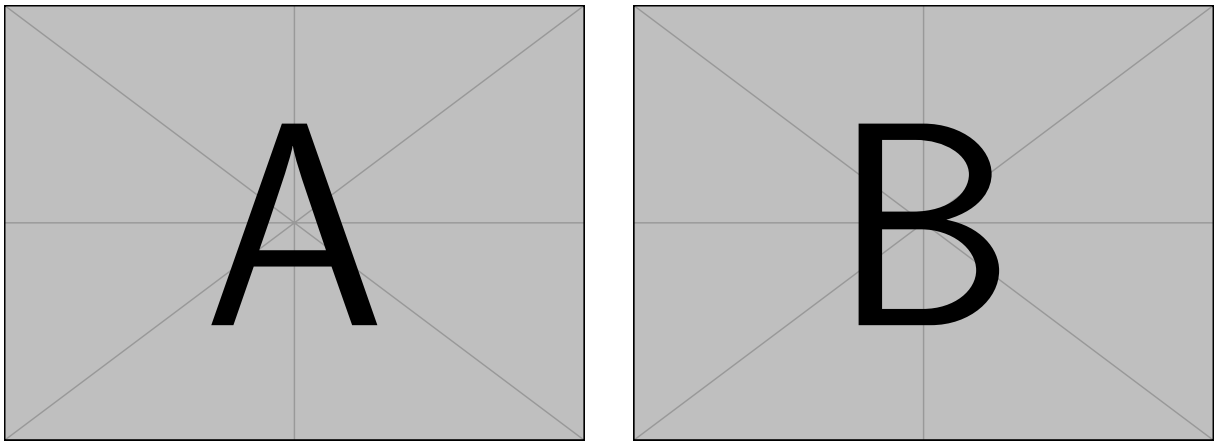

Figure 4: A minimal working example to demonstrate how to place two images side-by-side.

Output	natbib command	ACL only command
(Gusfield, 1997)	`\citep`	
Gusfield, 1997	`\citealp`	
Gusfield (1997)	`\citet`	
(1997)	`\citeyearpar`	
Gusfield's (1997)		`\citeposs`

Table 2: Citation commands supported by the style file. The style is based on the natbib package and supports all natbib citation commands. It also supports commands defined in previous ACL style files for compatibility.

2017 by Margaret Mitchell, ACL 2012 by Maggie Li and Michael White, ACL 2010 by Jing-Shin Chang and Philipp Koehn, ACL 2008 by Johanna D. Moore, Simone Teufel, James Allan, and Sadaoki Furui, ACL 2005 by Hwee Tou Ng and Kemal Oflazer, ACL 2002 by Eugene Charniak and Dekang Lin, and earlier ACL and EACL formats written by several people, including John Chen, Henry S. Thompson and Donald Walker. Additional elements were taken from the formatting instructions of the *International Joint Conference on Artificial Intelligence* and the *Conference on Computer Vision and Pattern Recognition*.

References

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

A Example Appendix

This is an appendix.