Carmen Easterwood

**Project Milestone**

* Evidence that you’ve been able to obtain, load, and play around a bit with your data. (For example, some simple exploratory data analysis.)
* Results from a baseline model. This can be very simple, such as random predictor, most-common-class, or a bag-of-words model.

1. **Abstract**

…

1. **Introduction (motivation for your work)**

…

GitHub repo: <https://github.com/carmen16/w266-final-project>

(Note specific code will be linked when it is discussed)

1. **Background (literature review, or related work)**

…

1. **Methods (include a description of any proposed work here, even if you haven’t done it yet)**

Full text, lead paragraph, nouns, lemmas

Table of # articles, vocab size, avg. unique words per article

Label cleaning

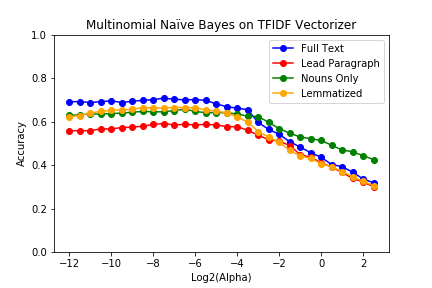
1. **Results and Discussion (for your baseline model, though feel free to include material for anything else you’ve done)**

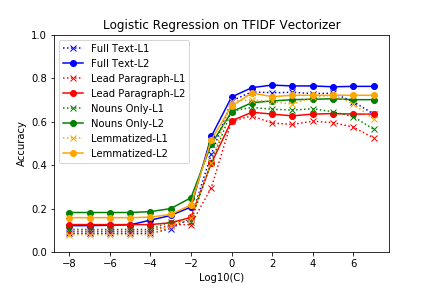
Full text wins both times

Best models (MNB: full text, alpha = 2^-7.5 = 0.00552, accuracy = 0.70703) (LR: full text, penalty = L2, C = 10^2 = 100, accuracy = 0.76757812)

Compare to models from references

Maybe lemmas/nouns better for unsupervised, but supervised benefits from as many words as possible





1. **Next Steps (section for work you plan to do before submitting the final version)**

Neural models

1. **References**

Fiona Martin and Mark Johnson. 2015. More Efficient Topic Modelling Through a Noun Only Approach. Proceedings of Australasian Language Technology Association Workshop, pages 111-115.

Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. 2017. Cross-Lingual Classification of Topics in Political Texts. Proceedings of the Second Workshop on Natural Language Processing and Computational Social Science, pages 42-46.

Mladen Karan, Jan Šnajder, Daniela Širinić, and Goran Glavaš. 2016. Analysis of Policy Agendas: Lessons Learned from Automatic Topic Classification of Croatian Political Texts. Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH), pages 12-21.

Stefan Wermter and Chihli Hung. 2002. Selforganizing Classification on the Reuters news corpus. Proceedings of the 19th international conference on Computational Linguistics – Volume 1.