Frameworks: Project Proposal

Documentation for the 'Frameworks' Project for FDAC 2018

Jerry Duncan
FDAC 2018 Student
The University of Tennessee
Knoxville, USA
jdunca51@vols.utk.edu

Trish Goedecke
FDAC 2018 Student
University of Tennessee Health Science Center
Memphis, USA
tgoedecke@uthsc.edu

Paul Preston Provins
FDAC 2018 Student
The University of Tennessee
Knoxville, USA
pprovins@vols.utk.edu

***Abstract*—We wish to explore the the predictive ability of topic models at various levels of dimensionality using correlations between topic prevalence and download rates to develop a model to predict popularity in JavaScript packages and frameworks.**

I. Motivation

Topic modeling is a technique for reducing dimensionality of large corpuses. Latent Dirichlet allocation (LDA), the most prevalent form of topic modeling, improves upon earlier methods by introducing Bayesian iterative updates, providing a sound theoretical basis for modeling by iteration. Yet a piece of the puzzle remains unsolved: how to select the most effective number of topics to model, K, is an unanswered question. This number of topics may also be considered the dimensionality of the model. We wish to explore the relationship between model quality and K, using both extrinsic measures (prediction) and internal metrics of the models.

II. Data to be Obtained

We will be using four different sources of data for this project.

- Github Commit History:
  Collect the projects that use these frameworks, the commits of the frameworks and the projects using them, and the authors of both.
- StackOverflow:
  Collect questions and answers which are related to each framework.
- NPM:
  Collect daily download count for each framework as well as projects that use them.
- Github Project Tracking:
  Collect Github issues and comments for each framework.

III. Usage of the Data

We plan to run LDA topic models on our full text data at varying levels of dimensionality 'K', from 10 to 100 topics. We may first expand the vocabulary using Wiki searches on a sample of our existing vocabulary.

We plan to break our text data and package download data into equal time segments (e.g. months). We will measure the topic prevalence and number of downloads per package in each time segment. For each level of K, we will correlate the topics with package download rates in one or more of the manners listed below. We will compare the effectiveness of each K's model in predicting download rates.

- Topic prevalence correlated with number of downloads during time segment
- Topic prevalence correlated with change in downloads during time segment
- Topic prevalence correlated with change in downloads during subsequent time segment
- Any of the above using rate of change of topic prevalence rather than raw prevalence per time segment

We will additionally run metrics on each K's model, which have been suggested as measures to optimize K in prior literature. We will measure perplexity (P), rate of change of perplexity (dP), and goodness of fit (GOF). We will compare our predictive findings with the findings across these mathematical metrics.

To complete the evaluation we will create an artificial corpus using Dirichlet allocation of a fixed number of topics to documents. We will run the metrics described above on our artificial corpus, to find which metric comes closest to selecting our predetermined level of K.

IV. Individual Group Member Responsibilities

Each individual in the group exhibit skills that they are proficient with, so work will be divided evenly and delegated based on each member's strengths. Our project consists of three main parts: collecting and cleaning data, formatting and feeding the data to topic models, and creating a website to query the data and interact with the results. Jerry is the most familiar with python libraries enabling the ability to collect,

clean, format, and collate data and will be in charge of writing the majority of data collection scripts that the rest of the group can reuse to get up-to-date data. Trish has a background in statistics and statistical modeling so she will be coordinating with Jerry to use the data in topic modeling and prediction. Preston is an expert in web development as well as database administration so he will be in charge of using the data Jerry collects to visualize the groups results and make the collected data queryable by visitors to the our website.

V. Milestones

Outlined below are the goals we have expected to accomplished at the corresponding date:

- September 26^{th}:
  By this point, the team will have finalized what data sources the team plans on using and what models will be used to analyze that data.

- October 10^{th}:
  Everyone on the team will have cloned the Frameworks repository and set up their development environment. Team meeting availability will be established, and the initial work will be delegated. The team will also become familiar with NPM, the selected frameworks, and topic modeling.

- October 24^{th}:
  Most, if not all, of the data that the team plans on using for modeling will have been collected and cleaned and in the process of being converted to a usable format for the topic modeling algorithms.

- November 7^{th}:
  The team will finish up the rest of the modeling and start trying to interpret the results and collate them into a comprehensible form for the final paper and presentation. Work on the website will begin so that the data and results gathered can be searched and used by the world.

- November 21^{st}:
  The final project paper and presentation slides will be completed and the website will be in a usable form in time for the presentation so that the rest of the class can see and visualize the team's data and results in real-time.

VI. Expected Outcome

Collectively we expect to have completed the tasks of collecting data from our proposed sources (Github, StackOverflow, and NPM). After data sanitization, we plan to have the data from select sources available in a queryable form to make our processes for analysis faster. From the collected data we will work to produce topic models at various levels of K which we anticipate will predict the popularity of a given framework. With the data collected and analysis made, we will utilize cloud computing and internet tool sets (Amazon Web Services) to provide our findings to anyone interested by publishing a website detailing our analysis with the queryable data we have collected.

Acknowledgment

Yuxing Ma of the University of Tennessee also contributed to this proposal and will have a limited role in this project.

References

[1] AlSumait, L., Barbara, D., Gentle, J., and Domeniconi, C. (2009). Topic significance ranking of LDA generative models. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 67-82. Springer Berlin Heidelberg.

[2] Bowman D. (2016). Selecting the number of topics in a latent Dirichlet allocation model. Statistical Learning and Data Science Section, Alexandria, VA, 1849-1857. American Statistical Association.

[3] Heinrich, G., Kindermann, J., Lauth, C., Paa, G., and Sanchez-Monzon, J. (2005). Investigating word correlation at different scopes–a latent concept approach. In Workshop Lexical Ontology Learning at Int. Conf. Mach. Learning.

[4] Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. (2010). Automatic evaluation of topic coherence. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 100-108. Association for Computational Linguistics.

[5] Stevens, K., Kegelmeyer, P., Andrzejewski, D., and Buttler, D. (2012). Exploring topic coherence over many models and many topics. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 952-961. Association for Computational Linguistics.

[6] Wallach, H. M., Murray, I., Salakhutdinov, R., and Mimno, D. (2009). Evaluation methods for topic models. In Proceedings of the 26th Annual International Conference on Machine Learning, 1105-1112. ACM.

[7] Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y., & Zou, W. (2015). A heuristic approach to determine an appropriate number of topics in topic modeling. BMC Bioinformatics, 16(13), S8.