Aidan Stoner, Miles Brandl, Jason Perrella

CSC 427 - Natural Language Processing

Project 2

Deliverable 3

- **What was easy about this assignment?**

  We found collaboration and communication between each team member easy. We often set up meet times working together to tackle relevant problems along the way. Each member took active effort to contribute and support each other. Overall group morale was high.

- **What was challenging about this assignment, or parts that you couldn't get working correctly?**

  Our group found our smoothed bigram models to be the most difficult part of the assignment. We originally had difficulty storing every bigram count in a dictionary since we didn't know how to allocate more ram in ELSA. As a result we attempted to create a sparse dictionary of bigram counts for every occurring bigram before smoothing, adding 1 to each unseen bigram on the spot when calculating probabilities. We found that ~99% of possible bigrams do not occur, so add-1 smoothing greatly shifts probability and creates continuous run-on sentences.

- **What did you like about this assignment?**

  We liked generating sentences once we got our bigram models and sentence generation working. Making coherent sentences consecutively generated was fun given that they were generated completely off of n-gram probabilities. Learning the fundamentals of n-gram predictive models was fascinating as we were able to implement them from scratch and see how probability and structure interact to form sentences.

- **What did you dislike about this assignment?**

  Some things that we didn't necessarily dislike but had some issues with were understanding smoothing techniques conceptually. While it made sense as we progressed, it was somewhat confusing implementing it and working with it. Also, sentence generation was relatively unpredictable at times, especially with the bigram model, which made debugging difficult. Lastly, integrating all the parts of the project into one unified interface took longer than expected, but we figured it all out in time.

- **How did your team function? Include details regarding what each team member contributed, how the team communicated with each other, and how team software development & design was accomplished.**

  T1 - Jason, Aidan, Miles : Implemented functions to compute unsmoothed unigram and bigram probability estimates using Maximum Likelihood Estimation.

  T2 - Aidan : Added functions to probabilistically generate sentences using our probability estimates.

  T3 - Miles : Added a function to apply add-1 smoothing to our language models.

  T4 - Miles, Aidan : Added a function to compute the perplexity of a test set.

  T5 - Jason, Aidan : Found two corpora and conducted analyses regarding the most probable uni,bi-grams and their probabilities for both corpora.

  T6 - Aidan, Miles : Developed functions to generate random sentences using our two language models. Also, developed analyses for them.

  T7 - Miles, Jason : Repeated T5 and T6 with Add-1 smoothing and wrote analyses comparing and contrasting the Add-1 models with the MLE models.

  T8 - Aidan, Miles : Computed and analyzed the perplexity measured out on random subsets of the corpora (held out test sets) for both smoothed models and corpora.

This is a brief overview of how our team tackled each task. We all worked collaboratively but we assigned ourselves with designated portions of the assignment to work on specifically. We communicated very consistently and often via iMessage and met up physically to work on the project together. We shared a code repository on GitHub and used several debugging methods to improve our model’s performance.

- **What did you learn from this assignment?**

  From this assignment, we learned several core NLP concepts and implemented them. For one, we furthered our understanding of n-gram language models, specifically uni- and bigram, and how they are computed using Maximum Likelihood Estimation(MLE). We also learned how to address issues with unseen data using the add-1 smoothing technique. We also gained hands-on experience with generating sentences using statistical language models, which helped us understand the randomness and limitations of these language models. Perplexity is another key concept we learned that provided us with a metric to determine how well our language model predicts new data. Lastly, we all sharpened our skills in terms of team collaboration, such as dividing tasks effectively, using consistent communication via iMessage group chat, and shared files through a common workspace using GitHub, helping us stay organized and accountable.