Project Type 2: N-gram Language Model

Sangmi Yun, Ulysses Galvez, Lindsay Hippe

1. Introduction

   For this n-gram assignment, we chose to compare two datasets of monolog read speech: one with dysfluent speech and one with fluent speech. For the preprocessing, we changed all the capital letters into lowercase letters and deleted extraneous characters, such as parentheses, numbers, and hyphens, in order for the result to represent solely the words that were spoken and whether they were spoken fluently. After preprocessing, we used Python to create our language models. We relied on the NLTK package to streamline the process. We drew heavily upon the code from Analytics Vidhya as we are all still very new to coding and do not yet have a sophisticated enough understanding of Python to complete this task from scratch on our own. We modified this source code in order to adequately fit our needs. Additionally, we were able to complete the preprocessing on our own as it felt like a much simpler task. Since our stuttering dataset was relatively small, we chose to use bigrams, which we imported from NLTK. Based on the input, we counted the frequency of the word, and then calculated the n-gram (probability) p(w|h) = w/h. After all the processing, we selected the most likely word from this probability and added this word to the sentence using the append method.

2. Data

   Dysfluent: UCLASS Release Two
   https://www.uclass.psychol.ucl.ac.uk/Release2/Monologue/FlatOrtho/index.htm
   https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2939977/

   The dysfluent dataset consists of four monolog recordings orthographically transcribed from three speakers. All speakers were school-aged males and spoke English as their primary language at home. We are using the second release of monologue datasets, from the stuttering dataset released by UCLASS in 2008. Two monolog recordings from different speakers have been phonetically transcribed. Two read recordings have been orthographically transcribed. These datasets are focusing on one language, which is English.

   Fluent: LibriSpeech Dataset
   http://shachi.org/resources/5070
   LibriSppech is a corpus including 1000 hours of 16kHz read English speech, collected by Vassil Panayotov. The fluent datasets include 2484 people's monolog recordings (1283 male speakers and 1201 female speakers). The

language of the datasets is all English and the language area is America. This data is from reading audiobooks from the LibriVox project. They transcribed the sound file into the text.

3. Result

   Stuttering Dataset

   favorite:

   favorite animal is is that uh its uh i was there um i ' s well and so thats so thats so kind of my end of loads of an airfield in them um i come to school is science and many games um yeah it ' s absolutely fantastic um i come home from it ' m living the car and and we drive there and uh i spent two with even with uh yosemite and sometimes in two weeks over the weather just you know cough the uh i watch the the summer holidays i watch the most difficult part is science and and and um so thats so thats that is appalling as well bad because it was there for nine hours um kind of kings ' s very interesting

   LibriSpeech Dataset

   happy:

   happy to the game now and i was sitting clean and are trustworthy and after a like girls slept soundly in a myth is in his bridle rein around invalids and inevitable symptom of the staff and dance with her sweetest little shoulders of soul and which of it looking for their suppers or handy solomon or untrustworthy before that was taken three leaps for their court was the appearance who then said he ' s corner her mind like to him before he began to force of humour and pleasure when young man and go at oldclough who have some would put his feet to day of ice one of you ' s advice was very well beloved by prodding him placed on which is a gascon a word of all the lord said i and see the harem if you not married frank good by illness had been master that could see you phronsie decidedly as soon as you the mouth right honourable mister frost ' s little man upon his learning and historiographer and it is that guards prided themselves by the same time and her with him objection one for our sovereign set for instance could steal on foot and unhappy as this thy maiden received for some question a spell he looked gravely in dogma it ' s but struggling life but that way you has the single woman to her name of socrates says the morning will defend us a special information which had copied from some money that officer and as if it but which some insect which had gone men who does speaking i didn ' s orders not let this expression of neck of them an effort and what it ' s bad as soon as well concede so grieved was do i didn ' t care of rhode island and listened at lelechka lelechka there is also to his eyes bleared and had voted to read the holy one matter drop of him she was merited my father who will and stretched out i surmise had every

log loaded on the idea of the way of scorpion sting of course he was apparently without avail that duke or duration for that bashful downcast face made such a murmur was without avail that geraint came to mister murry boldly faces that one obvious distinction being made him with them thus to him quite mistaken i have you can and day the painter appeared before he sang a brain upside down on his funeral upon every day had risen and pale with glory on at some little plans for me so let him to some good gait among fools and sent him to which everybody in his eyes to begin publishing that although they tell him a nervous bilious subject i couldn ' ve made her mother ' s green tent were they were written three dollars safe enough to flee i have waged war on the same to meet her sit on his sword for the galleries of loading i have told dorothy and we had to get a magnificent estate said griggs dryly they ' s horse quoth i went into a successful in what a sign the early she answered that makes me why the men had been answered him no one of the homage so many other cases i don ' s orders the king that was a silly bird dont you down to a god and that swift to henrietta but you lie he stood before him out into his feet to me you don ' clock i shall probably just recently

4. Discussion

    Compare the two models' performance. Consider calculating perplexity using a single test data for both models. Is one model more generalizable? How does training size impact performance? How does dataset scope impact performance? What sort of bias do these data encode?

    The model trained on fluent speech is more generalizable than the stuttering model because it has a larger training size. It includes data from a greater variety of speakers and has a greater number of utterances, therefore it can be applied to a greater array of the population. The training size certainly impacts performance in that a larger training size will likely improve performance. The scope of the dataset also impacts performance because it directly affects the topics the output of the model will address. For example, a model trained on data about dogs will not produce output concerning other topics such as plants. However, there is a great deal of overlap between various subjects so it is certainly possible that the output will mention things that are marginally related.

    As seen above, one of the biggest differences between the stuttering dataset and the fluent dataset is repeating words. With this, we can assume that the n-gram program is predicting the disfluencies that are present in the stuttering training data. This is quite fascinating because while humans are apt to ignore disfluencies as they obscure the meaning of the utterance, these probabilistic

models necessarily assign repeated words the same importance as unique words.

References

Bender, & Friedman, B. (2018). Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. Transactions of the Association for Computational\Linguistics, 6, 587–604. https://doi.org/10.1162/tacl_a_00041

Bird, Steven, Edward Loper and Ewan Klein (2009), *Natural Language Processing with Python.* O'Reilly Media Inc.

Howell, P., Davis, S., & Bartrip, J. (2009, April). *The University College London Archive of stuttered speech (UCLASS)*. Journal of speech, language, and hearing research : JSLHR. Retrieved June 5, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2939977/

Monolog flat orthographic format. Monoflatorth.(n.d.). Retrieved June 5, 2022, from https://www.uclass.psychol.ucl.ac.uk/Release2/Monologue/FlatOrtho/index.htm

NLP: Build Language Model in Python. Analytics Vidhya. (2020, December 23). Retrieved June 5, 2022, from https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-language-model-nlp-python-code/

Open Speech and Language Resources. Openslr.org. (n.d.). Retrieved June 5, 2022, from https://www.openslr.org/12