Project1 Part1 Report CS4740 Team members: Ransen Niu, Weicheng Yu

Problem Statement

Part 1 of this project uses unsmoothed unigram and bigram models to generate random sentences.

Preprocessing

We considered the email addresses at the beginning of the text be irrelevant text and removed them. And we removed the "bad" punctuations, which are usually irrelevant in generating sentences, including . Then we tokenized the text into an ordered list of words.

We consider all words are capital-sensitive. Also, we have a set of "stop" punctuations which indicate the end of the sentence.

Unsmoothed Unigram and Sentence Generation

We build the unigram model by taking counts of each word's frequency divided by the total number of the word tokens. To generate a sentence, we randomly sample words according to the unigram models. For example, a word w that has a unigram model P(w) = 0.1 would have 10 percent chance to be picked. When one of the "stop" punctuations appear, the sentence ends.

Sentence Examples

Sentence Analysis

Unsmoothed Bigram and Sentence Generation

For bigram model, we prepend a beginning marker "
begin" and append a stop marker "
stop" to the ordered list of words. Then we go through the element (word) in the list two by two to count the frequency of each pair of them. To generate a sentence, we first need to determine the start of the sentence. We choose the beginning word according to the unigram model and the capitalization of the word, i.e. we only consider words that have the first letter

Sentence Examples

Sentence Analysis

Comparisons between Unigram and Bigram models

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

Analysis & Testing

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Final Evaluation

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

Attachments

Lab Notes, HelloWorld.ic, FooBar.ic

References

[1] Fred G. Martin *Robotics Explorations: A Hands-On Introduction to Engineering.* New Jersey: Prentice Hall.

[2] Flueck, Alexander J. 2005. *ECE 100*[online]. Chicago: Illinois Institute of Technology, Electrical and Computer Engineering Department, 2005 [cited 30 August 2005]. Available from World Wide Web: (http://www.ece.iit.edu/ flueck/ece100).