CS 4740/5740 – Natural Language Processing

Fall 2017

Project 1 - Part 1 - Report

Team members: Ishaan Jain(irj4), Darpan Kalra(dk557), Tommy Shum(ts539)

Notes:

Our group decided to count punctuation marks as a word type and word token.

Sample Sentences:

Positive Unigram Sentences:

fight actors need the of strength . film ; stays
. sportsmen movies a Rachel the a the boy of
" , debut

Negative Unigram Sentences:

quirks rent creepiest only
unsettling is to action character-who-shall drugs plays book-on-tape great up
hairier it dreary and yawning office result enhance . sounds

Positive Bigram Sentences:

<s> The film about the same . </s>
<s> Worth a strong voices . </s>
<s> Does an engaging storyline , earnest dramaturgy , earnest , </s>

Negative Bigram Sentences:

<s> I found in neutral , it could have been in </s>
<s> Has nothing redeeming about three gags . </s>
<s> We 've been better . </s>

Analysis:

The sentences generated by the bigram model are more realistic and natural compared to those generated by the unigram model. The sentences generated by the unigram model have an unnatural sentence structure. For example, since our model counts punctuation as a word type, many unigram model sentences started off with punctuation. On the other hand, none of our bigram model sentences started with punctuation. In addition, the bigram model sentences conform better to the subject + verb + object pattern. Also, the bigram model sentences always correctly start with a capitalized word. This is not always the case for the unigram model sentences. On a related note, bigram model sentences generally end with appropriate punctuation. For example, many of the sample bigram model sentences ended with a period. In comparison, none of the sample sentences generated by the unigram model ended with appropriate punctuation.

The sentences generated by the bigram model generally convey positive and negative sentiments better that those generated by the unigram model. For example, the sample positive sentences generated by the unigram model are all very vague. It is not clear that these are positive reviews. On the other hand, the third positive review produced by the bigram model includes the phrase ‘Does an engaging storyline’, which clearly indicated that this is a positive review. Similarly, the negative reviews generated by the unigram model are vague and unintelligible. None of the three sample reviews generated by the unigram model are clearly negative. In contrast, the second negative review generated by the bigram model contains the phrase ‘Has nothing redeeming about’, which is clearly negative. Therefore, the bigram model sentences are more effective at conveying negative or positive opinions.