Automatic Summary Generation for Online Product Reviews

Kaveri Gupta(u1077655)
Sagar Chaturvedi(u1068847)

1: Problem Statement and Dataset

Our project aims to solve the problem by creating summaries of online product reviews while keeping the original sentiment.

We have used amazon reviews dataset available at `http://jmcauley.ucsd.edu/data/amazon/` for 8 categories: Beauty, call phones and accessories, clothing, shoes and jewellery, digital music, grocery and gourmet food, office products, sports and outdoor and toys and games.

2: Technical Design
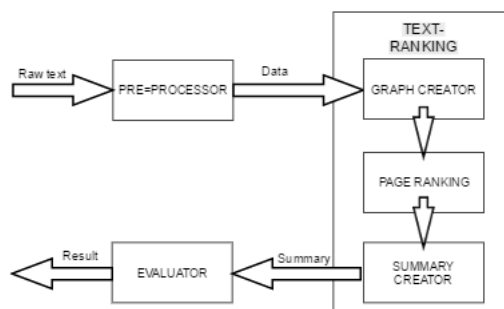
1. DATA PRE-PROCESSING:
   For pre-processing, we filtered out the reviews that did not meet the following criteria:

   (a) summary had less than 10 words or greater than 15 words.

   (b) review text was less than five times the summary text.

   (c) review was written before 2013.

   Note: For the later phase, we would like to filter data based on similarity between summary and review text to eliminate irrelevant summaries.

2. ARCHITECTURE

   (a) ALGORITHM:
       We have implemented text rank algorithm as discussed in the paper[1]. The steps are as follows:

       i. Tokenize the text. Here, one sentence in the review is one token as we're interested in importance of the sentence as whole and not individual words or phrases.

       ii. Build a graph as follows:

      A. For each token (sentence) create a node in the graph.

      B. Create an edge between all the nodes in the graph i.e. a fully connected graph.

      C. For each edge, compute the edit distance between the nodes (sentences) connected by the edge. Assign this distance as weight of the edge.

   iii. Page Rank algorithm is run on the generated graph.

   iv. The nodes/ sentences are then sorted based on the decreasing value assigned by the page rank algorithm.

   v. A summary of varied lengths is generated by concatinating the sentences based on their importance.

(b) EVALUATION METRICS:

We have used two metrics to evaluate the quality of the summaries generated by our system. They are:

   i. ROUGE:

We have used 1-gram, 2-gram, 3-gram and 4-gram models for evaluating recall, precision and f-score system summaries against the gold standard summaries.

   ii. BLEU:

We have used 4-gram BLEU model with '+1 smoothing' for evaluation.

3. RESOURCES USED:

Following is the list of Python libraries and external tools used:
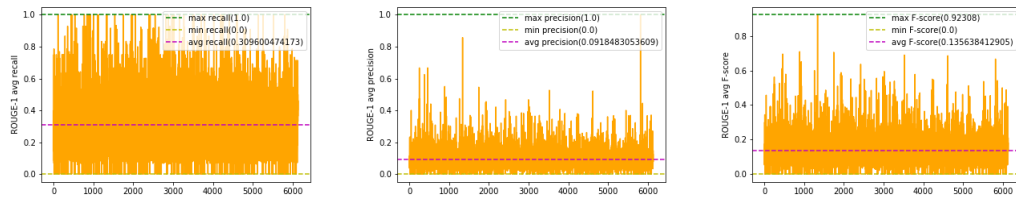
(a) *Python packages*: nltk, networkx, json, pandas, itertools, matplotlib

(b) *External tools*: Java implementation of ROUGE.
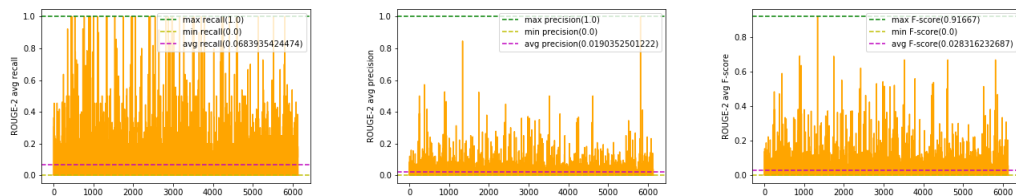

3: Evaluation Results:

We tested our system on 6137 instances and got the following results:
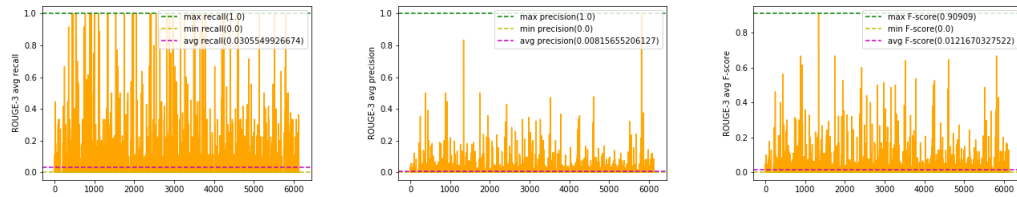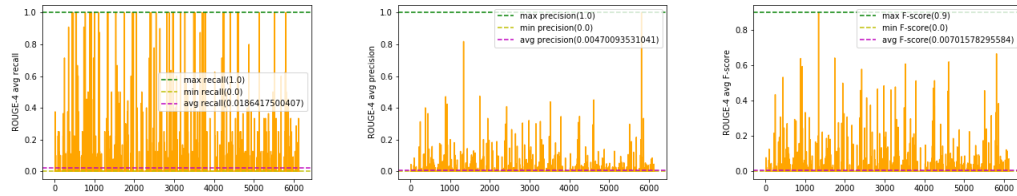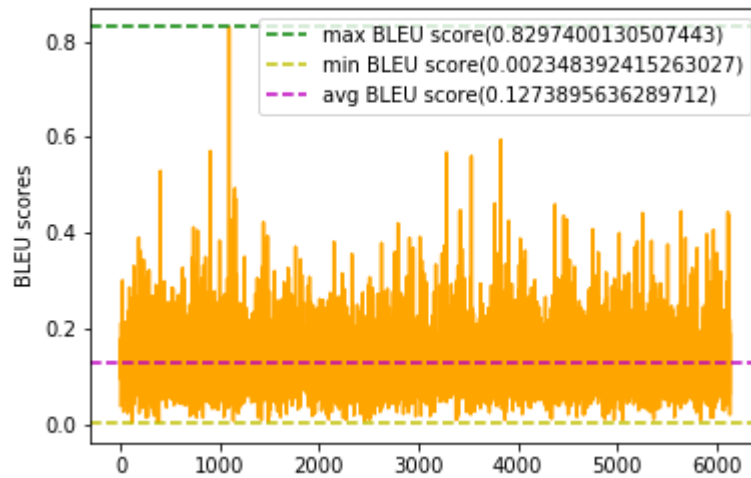
1. ROUGE:

(a) 1-gram

(b) 2-gram

(c) 3-gram

(d) 4-gram

2. BLEU:

4: Sample Output:

Following are the few summaries generated by the system:

1. BLEU SCORE: 0.8297400130507443
   SYSTEM: As today, this is the worst USB charging cable I ever seen and bought.
   REFERENCE: this is the worst USB charging cable I ever seen and bought

2. BLEU SCORE: 0.5943832348553569
   SYSTEM: Like my other Spyderco pocket knife it is well manufactured and was very sharp right out of the box.
   REFERENCE: Like my other Spyderco pocket knife it is well manufactured ...

3. BLEU SCORE: 0.003955857068927248
   SYSTEM: it's a powerful gun with excellent range, but it only lasted about 6 months because i left the CO_2 cartridges in the compartment for several months and after that, every time i put a new cartridge in with a minimal amount of gas leaking out when the needle punctured the top of it, it would not shoot, airsoft store employees told me something along the lines of it getting ruined because the effect of the metal piercing needle in the metal cartridge ruined the gun, so the needle was worn out which was exactly what it was, they recommended gas green-powered guns better than CO2
   REFERENCE: DO NOT LEAVE THE GAS CARTRIDGE IN FOR A LONG TIME

5: References

1. Rada Mihalcea and Paul Tarau. 2004. Text Rank: Bringing Order into Text`https://web.eecs.umich.edu/~mihalcea/papers/mihalcea.emnlp04.pdf`