Auto-summarizing Text

Swetha Kolalapudi
CO-FOUNDER, LOONYCORN

www.loonycorn.com

Overview

Auto-summarize text using a rule based model

Scrape websites for text data using BeautifulSoup

Use NLTK for munging text- tokenization, stopword removal etc

Auto-summarizing text

I was given this book to review for Amazon Vine and I have to say after all the disasters that have followed I am glad that I did not pay for this book.

When I originally read the book I thought it was better than the works that she's followed up Something Borrowed/Something Blue with but the majority of this reason was because she wasn't glorifying cheating in this novel. It is a quick read and a nice tale but nothing I would go out of my way to recommend to friends but not something I hated. It was a good beach fluffy read but the ending somewhat annoyed me.

Then Emily ruined the book - and her novels - forever for me by acting [edited] out [/edit] online.

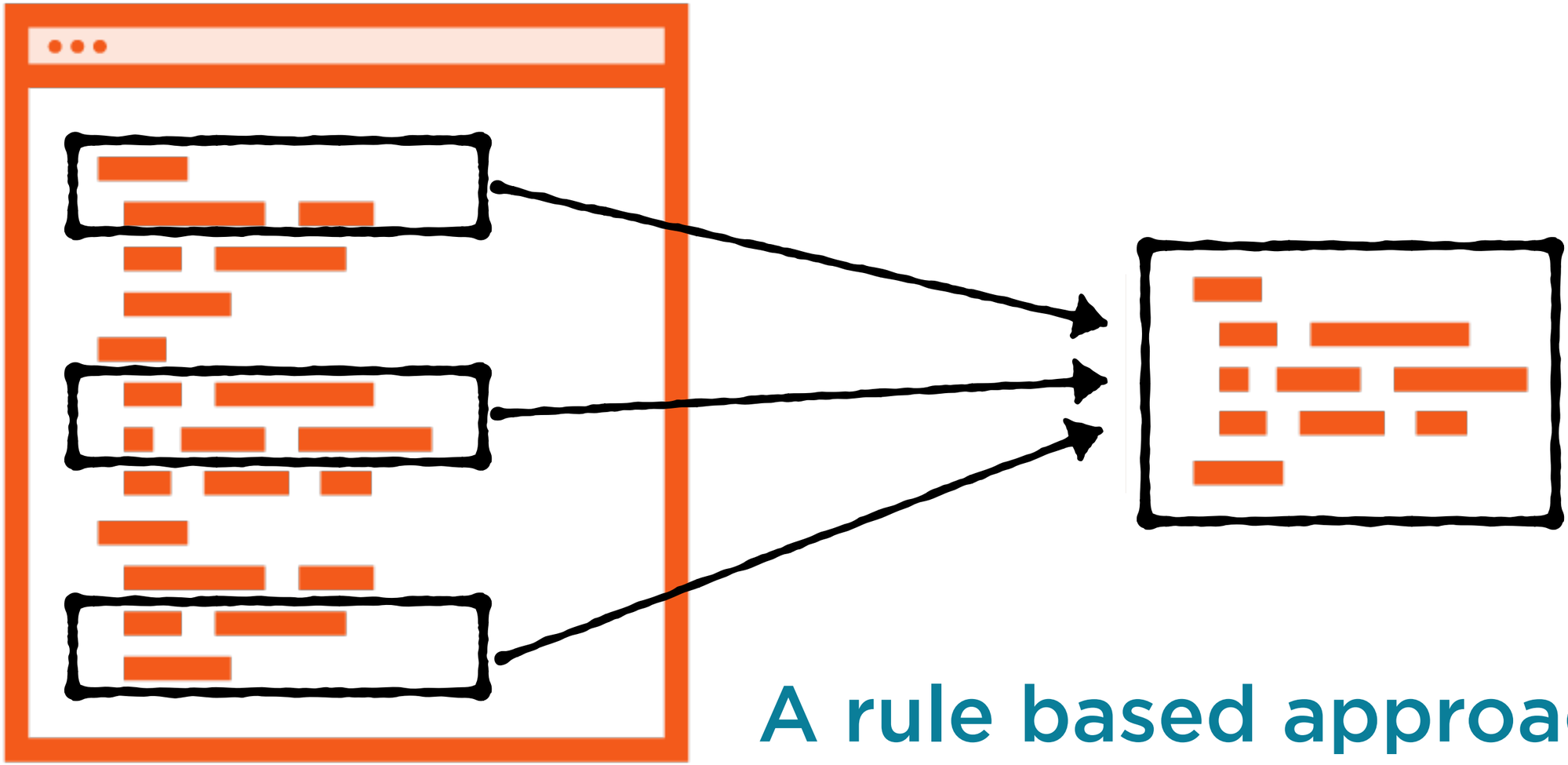
First she started off complaining excessively when her book didn't make #1 in the New York Times list and instead was number 2. She complained about this multiple times in multiple places, This was distressing to me to see as I know many authors would be thrilled to even BE on the NYT list much less in such an esteemed position. I couldn't believe how big of a deal she made of it.

Second her husband started attacking HONEST reviewers calling them psycho for not liking Emily's novel. Just wow. Is this even allowed under Amazon's rules of service? Then Emily proudly posted about it on her Facebook and Twitter and asked her fans to go and support her husband and post their own rebuttals (which turned into attacks) against the reviewer.

This has completely turned me against Emily Giffin and I will no longer read her novels. I may be one person but this is shameful.

“Emily ruined the book. This has completely turned me against Emily Giffin.”

Abstract Extraction



A rule based approach

Abstract Extraction

- Find the most important words
- Compute a significance score for sentences based on words they contain
- Pick the top most significant sentences

Abstract Extraction

- Find the **most important words**
- Compute a significance score for sentences based on words they contain
- Pick the top most significant sentences

Word Importance

- Authors tend to repeat the words that are important to the theme of the text

Word Importance = Word Frequency

Abstract Extraction

- Find the **most important words**
- Compute a significance score for sentences based on words they contain
- Pick the top most significant sentences

Abstract Extraction

- Find the most important words
- Compute a **significance score for sentences** based on words they contain
- Pick the top most significant sentences

Sentence Significance

- **Sentences which encapsulate more of the important words are more significant**

Significance Score = Sum(Word Importance)

Abstract Extraction

Retrieve Text

Download and parse the text from a webpage

Preprocess Text

Tokenize text and remove stopwords

Extract Sentences

Rank words and sentences

Retrieve Text

Download a webpage

Parse text using BeautifulSoup

Preprocess Text

- Tokenize text into sentences
- Tokenize sentences into words
- Remove stopwords

Extract Sentences

Compute frequencies of words

Compute significance scores of sentences

Rank sentences by their score

Pick top N sentences

Demo

Download and parse text from a webpage

Demo

Preprocess downloaded text

Demo

Auto-summarize text by extracting the most important sentences

Summary

Auto-summarize text using a rule based model

Scrape websites for text data using BeautifulSoup

Use NLTK for munging text- tokenization, stopword removal etc