TEXT SUMMARIZATION

THIS IS A PRETTY CHALLENGING PROBLEM,
SO WE WILL MAKE FULL USE OF THE AMAZING
LIBRARIES THAT PYTHON PROVIDES

OBJECTIVE: TAKE IN THE URL OF A NEWSPAPER ARTICLE (FROM THE WASHINGTON POST), AND AUTOMATICALLY SUMMARIZE IT IN 3 SENTENCES

HOW? USING SOMETHING CALLED

NATURAL LANGUAGE PROCESSING

1. DOWNLOAD THE CONTENTS OF THE URL

CAN BE TRICKY - THE USUAL APPROACH IS TO TRY SOMETHING CALLED REGULAR EXPRESSIONS BUT THAT'S NOT VERY RELIABLE

2. EXTRACT THE ARTICLE FROM ALL THE OTHER HTML THAT IS IN THE WEBPAGE

LUCKILY, PYTHON HAS THIS AMAZING TEXT SCRAPING LIBRARY CALLED

BEAUTIFUL SOUP

3. FIGURE OUT WHICH THE 3 MOST IMPORTANT SENTENCES IN THE ARTICLE ARE

USE A NATURAL LANGUAGE PROCESSING ALGORITHM, IMPLEMENTED USING PYTHON'S NLP LIBRARY CALLED

NLTK

HOW TO FIGURE OUT THE MOST IMPORTANT SENTENCE IN AN ARTICLE?

HERE IS ONE APPROACH..

FIND THE MOST COMMON WORDS IN THE ARTICLE

SUCH WORDS ARE CALLED STOPWORDS

(OF COURSE, WE'D NEED TO ELIMINATE 'THE', 'IS', ETC)

FIND THE SENTENCE IN WHICH THOSE MOST COMMON WORDS OCCUR MOST OFTEN

BOOM! THAT'S THE MOST IMPORTANT SENTENCE

1. DOWNLOAD THE ARTICLE FROM THE URL

2. GET RID OF EVERYTHING OTHER THAN THE ARTICLE ITSELF

USE BEAUTIFUL SOUP

USE NLTK

3. SPLIT THE ARTICLE INTO WORDS

4. ELIMINATE THE STOPWORDS ('THE','IS'..)

5. FIND HOW OFTEN EACH REMAINING WORD OCCURS

6. THE MORE COMMON A WORD, THE MORE IMPORTANT IT IS

7. FOR EACH SENTENCE FIND A SCORE OF HOW IMPORTANT THE WORDS IN THAT SENTENCE ARE

8. RANK THE SENTENCES BY THAT SCORE