과제 개요

- 이번 과제의 목표는 주어진 문서를 세 줄로 요약해주는 프로그램을 만드는 것입니다
- 문서 요약 알고리즘으로는 TextRank를 사용하게 되며, 이를 파이썬으로 구현하면 됩니다
- 참고 문헌
  - Page, Lawrence, et al. *The PageRank citation ranking: Bringing order to the web*. Stanford InfoLab, 1999.
  - Mihalcea, Rada, and Paul Tarau. "TextRank: Bringing order into texts." *Association for Computational Linguistics*, 2004.

과제 개요

- 문서 요약 알고리즘
  - abstraction-based: 새로운 문장을 만들어서 요약
  - extraction-based: 있는 문장 중에서 중요한걸 선택하여 요약
- TextRank는 extraction-based summarization입니다
  - 각 문장이 얼마나 중요한지를 계산
  - 중요한 문장들을 몇개 골라서 요약으로 만듦

Description [edit]

PageRank is a link analysis algorithm and it assigns a numerical weighting to each element of a hyperlinked set of documents, such as the World Wide Web, with the purpose of "measuring" its relative importance within the set. The algorithm may be applied to any collection of entities with reciprocal quotations and references. The numerical weight that it assigns to any given element *E* is referred to as the *PageRank of E* and denoted by $PR(E)$. Other factors like *Author Rank* can contribute to the importance of an entity.

Cartoon illustrating the basic principle of PageRank. The size of each face is proportional to the total size of the other faces which are pointing to it.

A PageRank results from a mathematical algorithm based on the webgraph, created by all World Wide Web pages as nodes and hyperlinks as edges, taking into consideration authority hubs such as cnn.com or usa.gov. The rank value indicates an importance of a particular page. A hyperlink to a page counts as a vote of support. The PageRank of a page is defined recursively and depends on the number and PageRank metric of all pages that link to it ("incoming links"). A page that is linked to by many pages with high PageRank receives a high rank itself.

Numerous academic papers concerning PageRank have been published since Page and Brin's original paper.[5] In practice, the PageRank concept may be vulnerable to manipulation. Research has been conducted into identifying falsely influenced PageRank rankings. The goal is to find an effective means of ignoring links from documents with falsely influenced PageRank.[6]

Other link-based ranking algorithms for Web pages include the HITS algorithm invented by Jon Kleinberg (used by Teoma and now Ask.com),the IBM CLEVER project, the TrustRank algorithm and the hummingbird algorithm.[*citation needed*]

1. PageRank is a link analysis algorithm and it assigns a numerical weighting to each element of a hyperlinked set of documents, such as the World Wide Web, with the purpose of "measuring" its relative importance within the set
2. The PageRank of a page is defined recursively and depends on the number and PageRank metric of all pages that link to it ("incoming links")
3. A page that is linked to by many pages with high PageRank receives a high rank itself.

배경 설명 - 문서 전처리

- TextRank를 적용하기 위해선 두 문장간의 유사도를 계산해야함
- 주어진 두 문장이, 서로 단어가 얼마나 겹치는지를 유사도로 사용
- 그러나 한국어의 경우 조사가 있기 때문에 '서울은'과 '서울에서'가 다른 단어로 취급됨
- 이를 방지하기 위해 형태소 분석을 하여 각 어근에 품사를 표시해주는, 품사 태깅 전처리 과정이 필요
  - 서울/NNG 은/JX
  - 서울/NNG 에서/JKM
- 이 전처리 과정이 모두 완료된 파일이 제공됨
  - 제공된 파일들을 input이라고 생각하고 구현하시면 됩니다

배경 설명 - 문서 전처리

- 제공되는 파일: 총 3개
  - 더미 데이터 (input-dummy.txt)
  - 전처리가 끝난 한국어 뉴스 기사 (input-korean.txt)
  - 전처리가 끝난 영어 뉴스 기사 (input-english.txt)
  - 파일의 한 줄이 반드시 한 문장이 되도록 제공됨

Ford Motor Co (F.N) abruptly named James Hackett as chief executive on Monday, responding to investors' growing unease about the U.S. automaker's slumping stock price and its ability to counter threats from longtime rivals and Silicon Valley.

Ford Chairman Bill Ford Jr., whose family effectively controls the U.S. No. 2 automaker, said he wanted Hackett to speed up decision-making and cut costs, but did not offer specifics on how the new CEO should change operations.

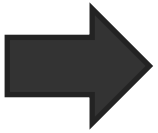
"The clock speed at which our competitors are working ...requires us to make decisions at a faster pace," said Ford Jr., who plans to take a more active role at the company, according to a person briefed on the matter.

Ford, which announced plans to cut 1,400 white-collar positions last week, is expected to look at further significant cost cuts in the next three to six months, according to company officials, speaking on condition of anonymity as the plans have not been finalized.

Hackett, 62, known as a turnaround expert who for the past year has led the Ford unit developing self-driving cars and related projects, replaces Mark Fields, 56, who spent less than three years as CEO.

Fields' abrupt dismissal caught nearly all at Ford by surprise, but concerns about the company's direction have been brewing for some time.

원본 기사

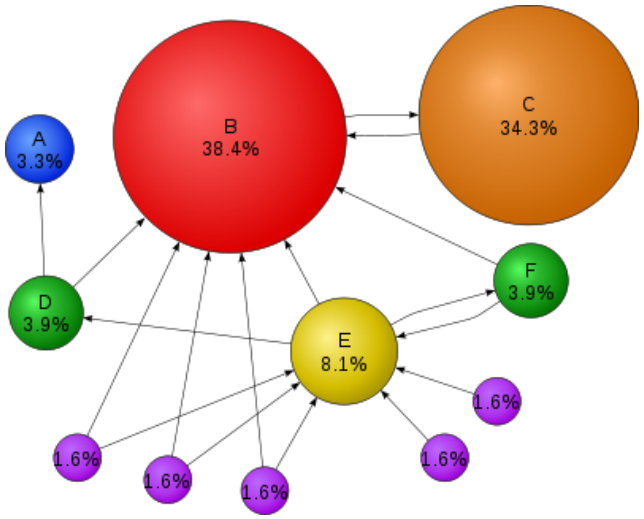


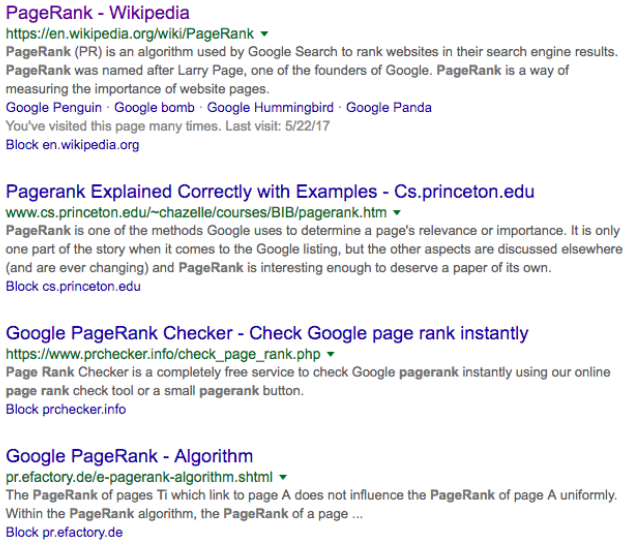
```
Ford/NNP Motor/NNP Co/NNP is/VBZ expected/VBN to/TO announce/VB the/DT departure/NN of/IN Chief/NNP
Executive/NNP Mark/NNP Fields/NNP in/IN a/DT broad/JJ management/NN shake-up/NN ,/, a/DT company/NN
source/NN said/VBD -/: a/DT move/NN that/WDT reflects/VBZ growing/VBG investor/NN unease/NN over/IN
the/DT company/NN 's/POS stock/NN market/NN performance/NN and/CC outlook/NN
Forbes/NNS and/CC the/DT New/NNP York/NNP Times/NNP reported/VBD that/IN James/NNP Hackett/NNP ,/, 62/CD
and/CC chairman/NN of/IN the/DT Ford/NNP unit/NN that/WDT works/VBZ on/IN autonomous/JJ vehicles/NNS ,/,
would/MD take/VB the/DT helm/NN
An/DT announcement/NN could/MD come/VB as/RB early/RB as/IN Monday/NNP
Ford/NNP shares/NNS are/VBP down/RB nearly/RB 40/CD percent/NN since/IN Fields/NNP ,/, 56/CD ,/,
took/VBD over/RP three/CD years/NNS ago/RB ,/, at/IN the/DT peak/NN of/IN the/DT U.S/NNP
auto/NN industry/NN 's/POS recovery/NN
Now/RB ,/, U.S/NNP
auto/NN sales/NNS are/VBP slipping/VBG ,/, and/CC Ford/NNP 's/POS profit/NN margins/NNS are/VBP
trailing/VBG those/DT of/IN larger/JJR rival/JJ General/NNP Motors/NNPS Co/NNP
Ford/NNP 's/POS board/NN of/IN directors/NNS and/CC Chairman/NNP Bill/NNP Ford/NNP Jr/NNP
have/VBP been/VBN unhappy/JJ with/IN the/DT company/NN 's/POS performance/NN ,/, and/CC sought/VBD
more/JJR reassurance/NN that/IN investments/NNS in/IN self-driving/JJ cars/NNS ,/, electric/JJ
vehicles/NNS and/CC ride/NN services/NNS would/MD pay/VB off/RP
Details/NNS of/IN further/JJ executive/NN moves/NNS were/VBD not/RB immediately/RB clear/JJ
The/DT Wall/NNP Street/NNP Journal/NNP reported/VBD on/IN Sunday/NNP that/IN the/DT company/NN was/VBD
considering/VBG new/JJ assignments/NNS for/IN some/DT of/IN Fields/NNP '/POS top/JJ lieutenants/NNS
``/`` We/PRP are/VBP staying/VBG focused/VBN on/IN our/PRP$ plan/NN for/IN creating/VBG value/NN and/CC
profitable/JJ growth/NN ,/, ''/'' a/DT Ford/NNP spokesman/NN in/IN Europe/NNP said/VBD in/IN response/NN
to/TO the/DT reports/NNS ,/, declining/VBG to/TO comment/VB ``/`` on/IN speculation/NN or/CC rumors/NNS
''/''
```

제공되는 파일

배경 설명 - PageRank

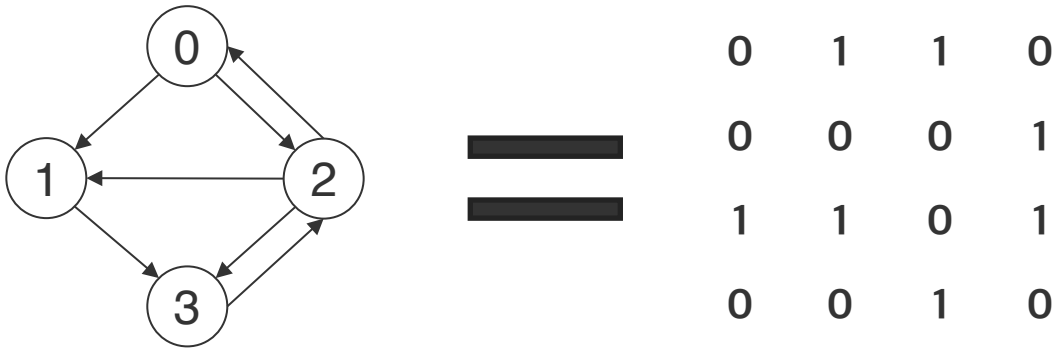
- 현재의 구글을 만든 핵심 알고리즘
  - “중요한 문서 A가 B를 참고하고 있다면, 문서 B 역시 중요할 것이다”
  - 문서(사이트)간 인용(링크) 정보를 그래프로 만들어서 중요도를 계산
  - 중요도 순으로 정렬한 결과가 바로 우리가 보는 구글 검색 결과 페이지
  - 추천 글: sungmooncho.com/2012/08/26/pagerank





배경 설명 - PageRank

- 문서간의 참조 관계를 directed graph로 표현
- directed graph를 행렬의 형태로 표현
  - i번 문서가 j를 참고했다면 a_{ij} = 1이 되도록 행렬 A = $\{a_{ij}\}$를 구성



  - 파이썬에선 A = [[0, 1, 1, 0], [0, 0, 0, 1], [1, 1, 0, 1], [0, 0, 1, 0]] 정도로 표현 가능

배경 설명 - PageRank

- 만든 행렬 A에서, 각 행의 합이 1이 되도록 normalize를 하고 transpose를 함

$$A = \begin{matrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{matrix} \Rightarrow \begin{matrix} 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 0 & 1 \\ 1/3 & 1/3 & 0 & 1/3 \\ 0 & 0 & 1 & 0 \end{matrix} \Rightarrow \begin{matrix} 0 & 0 & 1/3 & 0 \\ 1/2 & 0 & 1/3 & 0 \\ 1/2 & 0 & 0 & 1 \\ 0 & 1 & 1/3 & 0 \end{matrix} = M$$

- 그 결과를 행렬 M이라고 했을때
- 문서의 PageRank 열벡터 R은 다음 방정식의 해로 정의됨

$$R = dMR + \frac{1-d}{N}P$$

- N은 문서의 수, P는 원소값이 모두 1인 열벡터(Nx1행렬)
- d는 damping factor로써, 두번째 항과 함께 랜덤 점프를 나타냄
  - 보통 d = 0.85를 많이 씀

배경 설명 - PageRank

- 방정식을 풀면 다음과 같은 결과가 나옴

$$R = (I - dM)^{-1} \frac{1-d}{N} P$$

- 문서의 수가 많아질경우 역행렬 계산이 어렵기 때문에 iterative method로 R 값을 근사함

$$R(t+1) = dMR(t) + \frac{1-d}{N} P$$

- R(t + 1)과 R(t)의 차이가 충분히 작아질때까지 계산을 반복R은 열벡터(Nx1행렬), 따라서 '차이가 작다'는 '차이의 크기가 작다'로 해석
  - |R(t + 1) - R(t)| < e 가 될때까지
  - 본 과제에선 e = 0.000001로 설정
  - R(0)은 원소값이 모두 1/N인 열벡터

배경 설명 - PageRank

- 예시 (N = 4, d = 0.85, M은 앞선 예시, 수렴조건 0.000001)

$$M = \begin{matrix} 0 & 0 & 1/3 & 0 \\ 1/2 & 0 & 1/3 & 0 \\ 1/2 & 0 & 0 & 1 \\ 0 & 1 & 1/3 & 0 \end{matrix} \qquad R(0) = \begin{matrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{matrix} \qquad P = \begin{matrix} 1 \\ 1 \\ 1 \\ 1 \end{matrix}$$

$$R(t+1) = dMR(t) + \frac{1-d}{N}P$$

- t = 20에서 수렴
- R(20) = [0.1387, 0.1976, 0.3571, 0.3066] 정도 나오면 성공

배경 설명 - TextRank

- PageRank를 응용해서 만든 요약 알고리즘
- 기존 PageRank에서 행렬 A를 만드는 부분이 다름
- i번 문장과 j번 문장의 유사도가 k라면 a_{ij} = k가 되도록 행렬 A를 구성
  - 유사도는 대칭이므로 a_{ij} = a_{ji}
- A를 사용해서 M을 만들고, iterative로 R 수렴할때까지 계산하는 부분은 동일, d = 0.85, 수렴 조건 = 0.000001
- 유사도 함수로는 jaccard index를 사용 (4번 문제)

리마인드

- input.txt는 항상 한 줄이 한 문장이라고 가정해도 됩니다
  - 각 문장은 공백으로 단어가 구분되어 있습니다
- 요약된 결과는 원래 문장 순서대로입니다
  - 가령, 알고리즘의 결과상 1번-4번-0번, 순으로 중요하다고 계산이 되었다면, 0번-1번-4번 순으로 한줄씩 출력하셔야 합니다
- 3줄로 요약하는게 목표입니다
  - 문장의 중요도를 계산하고, 가장 중요한 문장 3개를 고르는것입니다
- PageRank 계산시 d = 0.85, 수렴 조건은 0.000001
- 뭔가 참고한게 있다면 보고서에 반드시 명시
  - 출처는 MLA나 APA 형식으로 기재하는것을 추천
- 5번의 경우, 실행 시키게되는 파일명만 summarize.py면 됩니다
  - 즉, 다른 파일들을 더 만드셔도 됩니다