Search Engines

Exercise 4: Text Processing

Dustin Lange & Saeedeh Momtazi
19 May 2011
Task 1: PageRank

• Apply the PageRank algorithm to our internet.
• Iteration 0: PageRank is equally distributed before the first iteration:
  \[ \text{PR}(A) = \text{PR}(B) = \text{PR}(C) = \text{PR}(D) = \frac{1}{4} \]
• With a random visit probability of \( \lambda = 0.2 \), compute the first four iterations and always round to three places after the comma. Show your calculations in detail for the first two iterations. Do the values converge?
Task 2: Stemming

• Select a German and an English text (from the texts you chose in the last exercise). Process both texts using two stemmers: The Porter stemmer and the German stemmer from Snowball (http://snowball.tartarus.org/).

• For each text, find 5 examples of differences in the stemming (with the different stemmers) that could have an impact on ranking.

• Give example queries for which you would expect different results. (Optional)
Task 3: Text Processing

• Text processing techniques have an important impact on information retrieval.

• Their contribution is not always apparent (improving the IR performance).

• They provide benefits that go beyond quantitative metrics to support broader tasks and richer experiences.

• This is a brainstorming task to find the role of text processing techniques in search engines.
Task 3: Text Processing

For each of the following techniques, give an example, a new idea, or a new challenge which shows how the technique can contribute in information retrieval:

- Language Modeling (N-grams)
- Machine Translation
- Named Entity Recognition
- Parsing
- Part of Speech Tagging
- Relationship Extraction
- Stemming
- Spell Correction
- Summarization
- Text Categorization
- Word Prediction
- Word Sense Disambiguation
Task 3: Text Processing

Example for Language Modeling (N-grams):

new companies in York

companies in New York
new companies in York

"distant" says senior company - Firms urged to back York 10K race ...Looking for a new career? Find a job in York and all around North Yorkshire...

Denise Stuart, of York Enterprise, is tasked with recruiting new employees...

2 Nov 2013 - Denne Stuart, of York Enterprise, is tasked with recruiting new employees...

Show more results from yorkpress.co.uk

York Water Company financial and business news, updates, and information from The New York Times and other leading providers.
tojias.nytimes.com → Business → Companies → York Water Company

News for new companies in York

MarketShifting, a New York-based company, has been... (3 related articles)
Amato said he had the idea for MarketShifting while running Show Media, a New York company that puts billboards on top of taxis.

Seacast Online - Portsmouth NH, York ME, Hampton NH, Exeter NH...
While I was in New York City last week, I went through the street where all the Korean restaurants are and tried a Korean Fried Chicken joint...

For any measure, New York's economy depends heavily on small businesses (6 related articles)

My Account - New York & Company
Nyandcompany.com - Find trendy, sexy, comfortable, affordable clothing online. Womens apparel, accessories and footwear. Clothing available in sizes 2-18.

List of Motor Vehicle Dealers (New and Used) Companies in North...
We have 36 Motor Vehicle Dealers (New and Used) Companies in North York, Canada. See contact info and background details for free.

Category: Companies based in New York City - Wikipedia, the free...
Submissions & Next Exercise

• Submissions:
  – Create slides to present your solution.
  – Send us your presentation
    • as PDF or PPT(X) or ODP:
      SearchEngines4[Name1][Name2].[pdf|ppt|pptx|odp]
    • via e-mail with subject: Search Engines 4
    • to dustin (dot) lange (at) hpi (dot) ...
    • until 08 June 2011, 5:00 pm

• On 09 June 2011: Be prepared to present your solution
  – English (or German)
  – Absent: Send us an e-mail in advance
Thanks for Listening

• Updates
  – Mailing list: searchengines2011 (at) hpi (…)
  – See website

• Questions
  – Via e-mail:
    • dustin (dot) lange (at) hpi (…)
    • saeedeh (dot) momtazi (at) hpi (…)
  – Office: A-1.6 / A-1.7