Search Engines

Exercise 3:
Fingerprints and Zipf

Dustin Lange & Saeedeh Momtazi
5 May 2011
Task 1: Fingerprinting

• Write a program to generate fingerprints (not simhash) for documents. Use the program to detect duplicates within the given data set.
• You can use any reasonable hash function for the words.
• Report on the quality of the detection. How does the detection quality vary with following different settings?
  – Different n in n-gram size (compare at least 2 different n)
  – Random n-gram selection vs. 0 mod p
  – Different p in 0 mod p (compare at least 2 different p)
  – Different matching thresholds (number of shared fingerprints, compare at least 2 different thresholds)
Measuring Detection Quality: Precision & Recall

- **False negatives**
- **True positives**
- **False positives**
- **True negatives**

Data set

- **True duplicates**
- **Declared duplicates**

**Precision** = \[rac{\text{True positives}}{\text{Declared duplicates}}\]

**Recall** = \[rac{\text{True positives}}{\text{True duplicates}}\]

**F-Measure** = \[rac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}\]
Data for Task 1: Reuters News Articles

• Reuters News Articles:
  – Articles published over the period of one year (8/20/1996 – 8/19/1997)
  – Popular classification data set

• 1025 articles (subset)
  925 original articles
  + 100 artificial duplicates

• Two files
  – Articles: 1 article per line (line breaks removed)
  – Correct matches: Line numbers of duplicate articles (indexed starting with line 1)
Task 2: Zipf’s Law

• Select two different texts from the website www.gutenberg.org that have both English and German versions
  – The texts should be from different centuries: The newer text should have been published 150 years after the first one at the earliest.

Search Engines Exercise – Dustin Lange and Saeedeh Momtazi
Task 2: Zipf’s Law

• Determine the frequencies of the words in each of the 4 texts (2 English and 2 German)
• Plot the frequencies of the words in all four texts against their rank (the result should be a diagram with 4 plots)
• Do the plots confirm the Zipf law?
• Compare the results
  – English texts vs. German texts
  – Older texts vs. newer texts
Task 2: Zipf’s Law

• Determine the frequencies of the words in each of the 4 texts (2 English and 2 German)
• Plot the frequencies of the words in all four texts against their rank (the result should be a diagram with 4 plots)
• Do the plots confirm the Zipf law?
• Compare the results
  – English texts vs. German texts
  – Older texts vs. newer texts

Search Engines Exercise – Dustin Lange and Saeedeh Momtazi
Task 2: Zipf’s Law

- Determine the frequencies of the words in each of the 4 texts (2 English and 2 German)
- Plot the frequencies of the words in all four texts against their rank (the result should be a diagram with 4 plots)
- Do the plots confirm the Zipf law?
- Compare the results
  - English texts vs. German texts
  - Older texts vs. newer texts

Search Engines Exercise – Dustin Lange and Saeedeh Momtazi
Submissions & Next Exercise

• Submissions:
  – Create slides to present your solution.
  – Send us your presentation
    • as PDF or PPT(X) or ODP: 
      \[\text{SearchEngines3[Name1][Name2].[pdf|ppt|pptx|odp]}\]
    • via e-mail with subject: Search Engines 3
    • to dustin (dot) lange (at) hpi (dot) ... 
    • until 18 May 2011, 5:00 pm

• On 19 May 2011: Be prepared to present your solution
  – English (or German)
  – Absent: Send me 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