Information Retrieval

Practice Session 9

November 16, 2016

Assignment I - Text Categorization

- Grading starts soon...
- If you have submitted your assignment, you should have received a confirmation from us
- Groups should have stabilized by now
- We assume the groups stay fixed
- In case of necessary changes in the groups, contact us and do NOT modify the spreadsheet

Assignment II - Retrieval System (high level)

- Build your own Scala search engine
- The search engine returns a list of the most relevant documents for a number of queries
- The returned documents are ranked based on relevance

Assignment II - Retrieval System

- Build a complete IR system that:
  - Passes through the document collection once
    (two passes are acceptable too)
  - builds an inverted index
  - handles multiple queries simultaneously,
  - offers multiple relevance models
    (at least one term-based model and one language model),
  - outputs top n results per query,
  - calculates global quality metrics (P, R, F_1, MAP)

Code Base

- Start from scratch,
- or try our TinyIR code base
- code for I/O, parsing, formatting
- Use it as it is, do not modify the code
- Download it again!

Document Collection

- Tipster dataset
- $\approx$ 100 K news stories
- stories published 1987 - 1992
- 40 training queries with binary relevance

Download Data

- documents
- topics (= queries)
- qrels (= relevance judgements)

Relevance Judgements

```
51 0 FR891103-0032 1
51 0 AP880301-0271 0
...
51 0 DOE1-07-0319 1
52 0 WSJ870129-0011 1
...
```

- Topic
- Static 0 (ignore)
- Document ID (ok with or without dash)
- Relevant $\in \{0,1\}$

Topics

<top>

<head> Tipster Topic Description

<num> Number: 051

<dom> Domain: International Economics

<title> Topic: Airbus Subsidies

<desc> Description:

Document will discuss government assistance to Airbus Industrie, or mention a trade dispute between Airbus and a U.S. aircraft producer over the issue of subsidies.

<smry> Summary:

Document will discuss government assistance to Airbus Industrie, or mention a trade dispute between Airbus and a U.S. aircraft producer over the issue of subsidies.

<narr> Narrative:

A relevant document will cite or discuss assistance to Airbus Industrie by the French, German, British or Spanish government(s), or will discuss a trade dispute between Airbus or the European governments and a U.S. aircraft producer, most likely Boeing Co. or McDonnell Douglas Corp., or the U.S. government, over federal subsidies to Airbus.

<con> Concept(s):

1. Airbus Industrie

2. European aircraft consortium, Messerschmitt-Boelkow-Blohm GmbH, British Aerospace PLC, Aerospatiale, Construcciones Aeronauticas S.A.

3. federal subsidies, government assistance, aid, loan, financing

4. trade dispute, trade controversy, trade tension

Documents


```
▼<DOC>
  <DOCNO>AP880212-0001</DOCNO>
  <FILEID>AP-NR-02-12-88 2344EST</FILEID>
  <FIRST>u i AM-Vietnam-Amnesty 02-12 0398</FIRST>
  <SECOND>AM-Vietnam-Amnesty,0411</SECOND>
 ▼<HEAD>
    Reports Former Saigon Officials Released from Re-education Camp
  </HEAD>
  <DATELINE>BANGKOK, Thailand (AP)</DATELINE>
 ▼<TEXT>
    More than 150 former officers of theoverthrown South Vietnamese government have been released
    from are-education camp after 13 years of detention, the official VietnamNews Agency reported
    Saturday. The report from Hanoi, monitored in Bangkok, did not givespecific figures, but said
    those freed Friday included anex-Cabinet minister, a deputy minister, 10 generals, 115field-
    grade officers and 25 chaplains. It quoted Col. Luu Van Ham, director of the Nam Ha camp south
    ofHanoi, as saying all 700 former South Vietnamese officials who hadbeen held at the camp now
    have been released. They were among 1,014 South Vietnamese who were to be releasedfrom re-
    education camps under an amnesty announced by the Communistgovernment to mark Tet, the lunar
    new year that begins Feb. 17. The Vietnam News Agency report said many foreign journalists
    anda delegation from the Australia-Vietnam Friendship Associationattended the Nam Ha release
    ceremony. It said Lt. Gen. Nguyen Vinh Nghi, former commander of SouthVietnam's Third Army
    Corps, and Col. Tran Duc Minh, former directorof the Army Infantry Officers School, expressed
    ``gratitude to thegovernment for its humane treatment in spite of the fact that mostof them
    (detainees) had committed heinous crimes against thecountry and people.'' The prisoners had
    been held without formal charges or trialsince North Vietnam defeated the U.S.-backed South
    Vietnamesegovernment in April 1975, ending the Vietnam War. Communist authorities had called
    the prisoners war criminals andsaid they had to learn how to become citizens of the new
    society. Small numbers had been released occasionally without publicitybut the government
    announced last year that 480 political prisonerswould be freed to mark National Day on Sept.
    2. On Thursday, Vice Minister of Information Phan Quang said 1,014would be released under the
    Tet amnesty. He reported a total of 150 prisoners would remain in the camps,which he said once
    held 100,000. ``Depending on their repentance, they will gradually be releasedwithin a short
    period of time,'' Quang said. He said many of the former inmates would return to theirfamilies
    in Ho Chi Minh City, formerly the South Vietnamese capitalof Saigon. The amnesties apparently
    are part of efforts by Communist Partychief Nguyen Van Linh to heal internal divisions and
    improveVietnam's image abroad.
  </TEXT>
</DOC>
```

Administrativa

- Due: Tue 13th Dec 2016, 11:59 PM
- Submit top-100 results for 10 test queries (handed out later),
- Runnable Scala code,
- Execution instructions (README),
- Pdf report formally explaining your models and system and reporting training-data performance (max. 2 pages)
- to any of the TAs
- subject: ir-practical-2016-2-[groupid]

Submission format

```
91 1 FR891103-0032
91 2 AP880301-0271
...
91 100 DOE1-07-0319
92 1 WSJ870129-0011
...
```

- topic id
- rank
- tipster document id
- separated by a single space
- one result per line
- two files per team
- file name: ranking-[t|l]-[id].run

Code optimization

- Pay attention to memory (around 4 GB RAM)
- Consider optimizing your code to run faster
- Compare how much speed-up you get when using an inverted index vs. parsing a document collection in a single-pass streaming fashion in terms of avg. running time per query
- Include these numbers in your report

Grading

- Formally correct retrieval models
- Creativity bonus for models/methods that we did not cover in class
- Test-set performance
- Correct calculation of performance metrics

Questions?