# Exercise #2

## Introduction to Knowledge Management, WS2016

**Camilo Thorne**

(based on slides by T. Szyler and A. Melo)

Mannheim, 07.03.2016

# Exercise 02 – Question 1.1

## Task

Calculate the **idf** parameter for each term $i$:

$$idf_i = \log_{10}\left(\frac{N}{df_i}\right)$$

Do **not** use stemming or stopword removal. As shown in the lecture, the result of idf is high, if the term $i$ (which is one of our 8 words) occurs only in very few documents.

# Exercise 02 – Question 1.1

## Task

[$d_1$] **Play Games at Online Games, Games from Shockwave.com**
Play Games on the One-and-Only Shockwave.com. Shockwave.com is the ultimate destination for free online games, download games, and more!

[$d_2$] **Free Games Downloads + Free Online Games**
Free games downloads and free online games. Download free full version PC games, play free online games, arcade games and puzzles. Fun and friendly community

[$d_3$] **Free PC Games Download - Full Version PC Games**
Download Free PC Games Download - Full Version PC Games Download. Download Free PC Games. More Free Games. FREE GAMES HOME DOWNLOAD FREE GAMES ONLINE FREE

[$d_4$] **Download Free Games - Awesome Selection of Safe Game Downloads**
Download Free Games has over 1000 high quality and safe game downloads and free online games. Find a fun game, download free trials, watch game video

# Exercise 02 – Question 1.1

## TF and IDF

| w | D1 | D2 | D3 | D4 | df | idf |
|---|---|---|---|---|---|---|
| **PC** | 0 | 1 | 5 | 0 | 2 | 0.30103 |
| **Games** | 6 | 7 | 8 | 3 | 4 | 0 |
| **Free** | 1 | 6 | 7 | 4 | 4 | 0 |
| **Download** | 1 | 1 | 6 | 3 | 4 | 0 |
| **More** | 1 | 0 | 1 | 0 | 2 | 0.30103 |
| **Full** | 0 | 1 | 2 | 0 | 2 | 0.30103 |
| **Shockwave** | 3 | 0 | 0 | 0 | 1 | 0.60206 |
| **Quality** | 0 | 0 | 0 | 1 | 1 | 0.60206 |

$$idf_i = \log_{10}\left(\frac{N}{df_i}\right)$$

$$N = 4$$

# Exercise 02 – Question 1.1

## TF and IDF

| w | D1 | D2 | D3 | D4 | df | idf |
|---|---|---|---|---|---|---|
| **PC** | 0 | 1 | 5 | 0 | 2 | 0.30103 |
| **Games** | 6 | 7 | 8 | 3 | 4 | 0 |
| **Free** | 1 | 6 | 7 | 4 | 4 | 0 |
| **Download** | 1 | 1 | 6 | 3 | 4 | 0 |
| **More** | 1 | 0 | 1 | 0 | 2 | 0.30103 |
| **Full** | 0 | 1 | 2 | 0 | 2 | 0.30103 |
| **Shockwave** | 3 | 0 | 0 | 0 | 1 | 0.60206 |
| **Quality** | 0 | 0 | 0 | 1 | 1 | 0.60206 |

$$idf_i = \log_{10}\left(\frac{N}{df_i}\right)$$

$$N = 4$$

- queries are case insensitive
- **games**, **free** and **download** can be ignored

# Exercise 02 – Question 1.1

$w_{i,j}$

| | D1 | D2 | D3 | D4 | idf |
|---|---|---|---|---|---|
| **PC** | 0 | 1 | 5 | 0 | 0.30103 |
| **More** | 1 | 0 | 1 | 0 | 0.30103 |
| **Full** | 0 | 1 | 2 | 0 | 0.30103 |
| **Shockwave** | 3 | 0 | 0 | 0 | 0.60206 |
| **Quality** | 0 | 0 | 0 | 1 | 0.60206 |

$$w_{i,j} = tf_{i,j} \times idf_i$$

$\Rightarrow$ for every word $i$ and every document $j$:

| | D1 | D2 | D3 | D4 |
|---|---|---|---|---|
| **PC** | 0 | 0.30103 | 1.50515 | 0 |
| **More** | 0.30103 | 0 | 0.30103 | 0 |
| **Full** | 0 | 0.30103 | 0.60206 | 0 |
| **Shockwave** | 1.80618 | 0 | 0 | 0 |
| **Quality** | 0 | 0 | 0 | 0.60206 |

# Exercise 02 – Question 1.2

## Cosine Similarity

$$sim(q_j, d_k) = \frac{q_j \bullet d_k}{||q_j|| \times ||d_k||} = \frac{\sum_i q_{i,j} \times w_{i,k}}{\sqrt{\sum_i q_{i,j}^2} \times \sqrt{\sum_i w_{i,k}^2}}$$

where

1. $q_{i,j} = \begin{cases} 1, & \text{if the word } i \text{ is contained in query } j \\ 0, & \text{otherwise} \end{cases}$
2. $d_k$ = vector representation of document k (each of the 4 documents)
3. $q_j$ = vector representation of query j (each of the 2 queries)
4. $||\cdot||$ is the Euclidian norm and $\bullet(\cdot,\cdot)$ the vector dot product

# Exercise 02 – Question 1.2

## Cosine Similarity - Computation

Given $q_j$, to compute $sim(q_j, d_k)$ for each $d_k$
one usually proceeds in steps:

1. compute $q_j$ (binary –$0/1$– vector in $|W|$ space)
2. **compute the norm** $||q_j||$ of $q_j$
3. compute $\sum_i q_{i,j}$
4. for each document $d_k$ do:
   - **compute the norm** $||d_k||$
   - for each term/word $i$ occurring in the query $q_j$, compute $w_{i,k}$
5. plugin the quantities in the $sim(q_j, d_k)$ formula and **return the similarity score**

# Exercise 02 – Question 1.2

## Calculate Denominator, $q_1$: PC games free download

$\sqrt{\sum_i q_{i,j}^2}$:

$$
\begin{aligned}
q_{PC,j} &= 1 \\
q_{games,j} &= 1 \\
q_{free,j} &= 1 \\
q_{download,j} &= 1
\end{aligned}
$$

hence:

$$\sqrt{\sum_i q_{i,j}^2} = \sqrt{1^2 + 1^2 + 1^2 + 1^2} = \sqrt{4}$$

# Exercise 02 – Question 1.2

## Calculate Denominator, $q_1$: PC games free download

$\sqrt{\sum_i w_{i,k}^2}$:

| | D1 | D2 | D3 | D4 |
|---|---|---|---|---|
| **PC** | 0 | 0.30103 | 1.50515 | 0 |
| **More** | 0.30103 | 0 | 0.30103 | 0 |
| **Full** | 0 | 0.30103 | 0.60206 | 0 |
| **Shockwave** | 1.80618 | 0 | 0 | 0 |
| **Quality** | 0 | 0 | 0 | 0.60206 |

hence:

for $k = D1$ : $\sqrt{0.30103^2 + 1.80618^2}$
for $k = D2$ : $\sqrt{0.30103^2 + 0.30103^2}$
for $k = D3$ : $\sqrt{1.50515^2 + 0.30103^2 + 0.60206^2}$
for $k = D4$ : $\sqrt{0.60206^2}$

# Exercise 02 – Question 1.2

## Cosine Similarity, $q_1$: PC games free download

The words *games, free and download* are ignored as argued earlier. Thus, for query $q_1$ only **PC** causes any difference between the entries.

| $w_{i,j}$ | **D1** | **D2** | **D3** | **D4** |
|---|---|---|---|---|
| **PC** | 0 | 0.30103 | 1.50515 | 0 |

$$sim(q_1, d_1) = \frac{0}{\sqrt{4} \times \sqrt{0.30103^2 + 1.80618^2}} = \frac{0}{2 \times 1.83109} = \mathbf{0}$$

$$sim(q_1, d_2) = \frac{0.30103}{\sqrt{4} \times \sqrt{0.30103^2 + 0.30103^2}} = \frac{0.30103}{2 \times 0.42572} =$$

$$\mathbf{0.35355}$$

$$sim(q_1, d_3) = \frac{1.50515}{\sqrt{4} \times \sqrt{1.50515^2 + 0.30103^2 + 0.60206^2}} = \frac{1.50515}{2 \times 1.64881} =$$

$$\mathbf{0.45644}$$

$$sim(q_1, d_4) = \frac{0}{\sqrt{4} \times \sqrt{0.60206^2}} = \mathbf{0}$$

# Exercise 02 – Question 1.2

## Ranking, $q_1$: PC games free download

**Ranking** $q_1$

1. [**1.**] $d_3(0.45644)$
2. [**2.**] $d_2(0.35355)$
3. [**3.**] $d_1, d_4(0)$

# Exercise 02 – Question 1.2

## Calculate Denominator, $q_2$: Shockwave quality games

$$\sqrt{\sum_i q_{i,j}^2} = \sqrt{3}$$

We have three words in the query and no word occurs twice.

$\sqrt{\sum_i w_{i,k}^2}$:

Query independent value:

for $k = D1$ : $\sqrt{0.30103^2 + 1.80618^2}$
for $k = D2$ : $\sqrt{0.30103^2 + 0.30103^2}$
for $k = D3$ : $\sqrt{1.50515^2 + 0.30103^2 + 0.60206^2}$
for $k = D4$ : $\sqrt{0.60206^2}$

# Exercise 02 – Question 1.2

## Cosine Similarity, $q_2$: Shockwave quality games

Ignore **games** $\Rightarrow$ the words **Shockwave** and **quality** are the only important ones for the denominator:

| | D1 | D2 | D3 | D4 |
|---|---|---|---|---|
| **Shockwave** | 1.80618 | 0 | 0 | 0 |
| **Quality** | 0 | 0 | 0 | 0.60206 |

$$sim(q_2, d_1) = \frac{1.80618}{\sqrt{3} \times \sqrt{0.30103^2 + 1.80618^2}} = \frac{1.80618}{1.73205 \times 1.83109} = \mathbf{0.5695}$$

$$sim(q_2, d_2) = sim(q_2, d_3) = \mathbf{0}$$

$$sim(q_2, d_4) = \frac{0.60206}{\sqrt{3} \times \sqrt{0.60206^2}} = \frac{0.60206}{1.73205 \times 0.60206} = \mathbf{0.57735}$$

# Exercise 02 – Question 1.2

## Ranking, $q_2$: Shockwave quality games

**Ranking $q_2$**

1. [**1.**] $d_4(0.57735)$
2. [**2.**] $d_1(0.5695)$
3. [**3.**] $d_2, d_3(0)$

# Exercise 02 - Question 2

| rank | $a_1$ | $a_2$ | | $a_3$ | |
|---|---|---|---|---|---|
| 1 10 | $d_1$ | $d_1$ | $d_{11}$ | $d_1$ | $d_{15}$ |
| 2 11 | $d_2$ | $d_2$ | $d_{12}$ | $d_2$ | $d_{20}$ |
| 3 12 | $d_5$ | $d_4$ | $d_{13}$ | $d_4$ | |
| 4 13 | $d_6$ | $d_5$ | $d_{14}$ | $d_5$ | |
| 5 14 | $d_{13}$ | $d_6$ | $d_{15}$ | $d_9$ | |
| 6 15 | | $d_7$ | $d_{18}$ | $d_{10}$ | |
| 7 16 | | $d_8$ | $d_{20}$ | $d_{12}$ | |
| 8 | | $d_9$ | | $d_{13}$ | |
| 9 | | $d_{10}$ | | $d_{14}$ | |

Consider a search space $\mathcal{D}$ of 20 documents. For the query $q$ the first 10 documents $\{d_1, d_2, \ldots, d_{10}\}$ are **relevant**. Calculate the **(1) precision**, **(2) recall** and **(3) balanced F-measure** ($\beta = 1$ or $\alpha = 0.5$) values for each search algorithm $a_1$, $a_2$, and $a_3$ on $\mathcal{D}$.

# Evaluation Metrics

Reminder

## Precision

**Precision** is the number of correct results divided by the number of all returned results.

## Recall

**Recall** is the number of correct results divided by the number of results that should have been returned.

## F1-Measure

A measure that combines precision and recall is the **harmonic mean** of precision and recall, the traditional **F1-measure** or balanced F1-score.

cf: `http://en.wikipedia.org/wiki/Precision_and_recall`

# Evaluation Metrics

## Reminder

**Assume that 50 of 100 possible documents are relevant. Which precision and recall values result from an algorithm that...**

- ... returns all correct and no incorrect results?

# Evaluation Metrics

## Reminder

**Assume that 50 of 100 possible documents are relevant. Which precision and recall values result from an algorithm that...**

- ... returns all correct and no incorrect results? (Precision 1, Recall 1, F-Measure 1)

# Evaluation Metrics

## Reminder

**Assume that 50 of 100 possible documents are relevant. Which precision and recall values result from an algorithm that...**

- ... returns all correct and no incorrect results? (Precision 1, Recall 1, F-Measure 1)
- ... returns only incorrect and no correct results?

# Evaluation Metrics

## Reminder

**Assume that 50 of 100 possible documents are relevant. Which precision and recall values result from an algorithm that...**

- ... returns all correct and no incorrect results? (Precision 1, Recall 1, F-Measure 1)
- ... returns only incorrect and no correct results? (Precision 0, Recall 0, F-Measure 0)

# Evaluation Metrics

## Reminder

**Assume that 50 of 100 possible documents are relevant. Which precision and recall values result from an algorithm that...**

- ... returns all correct and no incorrect results? (Precision 1, Recall 1, F-Measure 1)
- ... returns only incorrect and no correct results? (Precision 0, Recall 0, F-Measure 0)
- ... returns only one correct result?

# Evaluation Metrics

## Reminder

**Assume that 50 of 100 possible documents are relevant. Which precision and recall values result from an algorithm that...**

- ... returns all correct and no incorrect results? (Precision 1, Recall 1, F-Measure 1)
- ... returns only incorrect and no correct results? (Precision 0, Recall 0, F-Measure 0)
- ... returns only one correct result? (Precision 1, Recall $\frac{1}{50}$, F-Measure 0.039)

# Evaluation Metrics

## Reminder

**Assume that 50 of 100 possible documents are relevant. Which precision and recall values result from an algorithm that...**

- ... returns all correct and no incorrect results? (Precision 1, Recall 1, F-Measure 1)
- ... returns only incorrect and no correct results? (Precision 0, Recall 0, F-Measure 0)
- ... returns only one correct result? (Precision 1, Recall $\frac{1}{50}$, F-Measure 0.039)
- ... returns all documents?

# Evaluation Metrics

## Reminder

**Assume that 50 of 100 possible documents are relevant. Which precision and recall values result from an algorithm that...**

- ... returns all correct and no incorrect results? (Precision 1, Recall 1, F-Measure 1)
- ... returns only incorrect and no correct results? (Precision 0, Recall 0, F-Measure 0)
- ... returns only one correct result? (Precision 1, Recall $\frac{1}{50}$, F-Measure 0.039)
- ... returns all documents? (Precision $\frac{50}{100}$, Recall 1, F-Measure 0.66667)

# Exercise 02 - Question 2.1

## Evaluation

Evaluation results for rankings $a_1, a_2$ and $a_3$

| | **Precision** | **Recall** | **F-Measure** |
|---|---|---|---|
| $a_1$ | | | |

# Exercise 02 - Question 2.1

## Evaluation

Evaluation results for rankings $a_1, a_2$ and $a_3$

| | **Precision** | **Recall** | **F-Measure** |
|---|---|---|---|
| $a_1$ | $\frac{4}{5} = 0.8$ | | |

# Exercise 02 - Question 2.1

## Evaluation

Evaluation results for rankings $a_1, a_2$ and $a_3$

| | **Precision** | **Recall** | **F-Measure** |
|---|---|---|---|
| $a_1$ | $\frac{4}{5} = 0.8$ | $\frac{4}{10} = 0.4$ | |

# Exercise 02 - Question 2.1

## Evaluation

Evaluation results for rankings $a_1, a_2$ and $a_3$

| | **Precision** | **Recall** | **F-Measure** |
|---|---|---|---|
| $a_1$ | $\frac{4}{5} = 0.8$ | $\frac{4}{10} = 0.4$ | $\frac{2 \times 0.8 \times 0.4}{0.8+0.4} = 0.53$ |

# Exercise 02 - Question 2.1

## Evaluation

Evaluation results for rankings $a_1, a_2$ and $a_3$

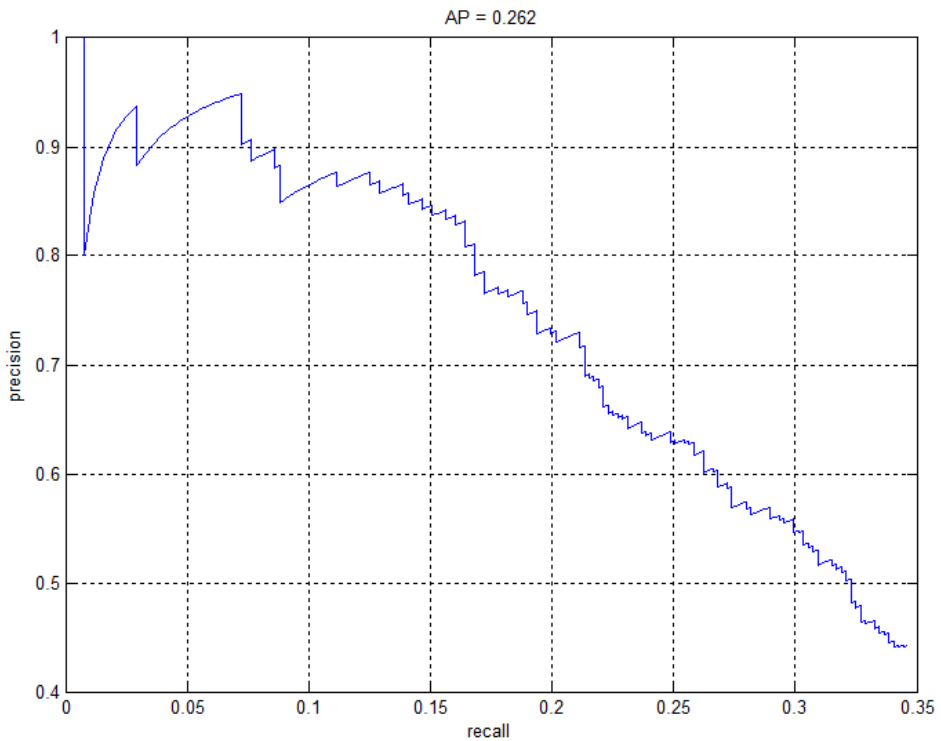
| | **Precision** | **Recall** | **F-Measure** |
|---|---|---|---|
| $a_1$ | $\frac{4}{5} = 0.8$ | $\frac{4}{10} = 0.4$ | $\frac{2 \times 0.8 \times 0.4}{0.8+0.4} = 0.53$ |
| $a_2$ | $\frac{9}{16} = 0.56$ | $\frac{9}{10} = 0.9$ | $\frac{2 \times 0.56 \times 0.9}{0.56+0.9} = \mathbf{0.69}$ |
| $a_3$ | $\frac{6}{11} = 0.55$ | $\frac{6}{10} = 0.6$ | $\frac{2 \times 0.55 \times 0.6}{0.55+0.6} = 0.57$ |

# Exercise 02 - Question 2.2

## Evaluation

Trade-off between precision and recall?

- Typically there's an inverse relationship between Precision and Recall



# Exercise 02 – Question 3.1a

## HelloLucene

Try out the following queries and report the output:

# Exercise 02 – Question 3.1a

## HelloLucene

Try out the following queries and report the output:

- Lucene
  1. *193398817 Lucene in Action*
  2. *55320055Z Lucene for Dummies*
- Action
  1. *193398817 Lucene in Action*
- lucene in action
  1. *193398817 Lucene in Action*
  2. *55320055Z Lucene for Dummies*

# Exercise 02 – Question 3.1a

## HelloLucene

- GigaBytes

  *1. 55063554A Managing Gigabytes*

- gigabyte

  *Found 0 hits.*

# Exercise 02 – Question 3.1b

## HelloLucene

- Why wasn't "Managing Gigabytes" found when you searched for "gigabyte"?
- Assume that the desired result of the query "gigabyte" is "Managing Gigabytes". Which type of linguistic preprocessing would have to be performed on the book titles, in order to get this result?

# Exercise 02 – Question 3.2

## HelloLucene

The search behaviour can be altered with special query operators. Compare the query pairs below: report their output and explain the differences.

- *lucene and action* vs *lucene AND action*
- *managuing* vs *managuing*$\sim$
- *a* vs *a**
- *"lucene NOT action"* vs *"lucene" NOT "action"*

# Exercise 02 – Question 3.3

## HelloLucene

Change the code so that the search works for the *isbn* field! Write down the resulting line of code and one example query along with its output.

# Hands on

# Alternative way to calculate Cosine Similarity & Uniterpolated Average Precision

(Not required for the exam)

# Calculate Denominator (Alternative way)

$q_1$: PC games free download

$\sqrt{\sum_i q_{i,j}^2}$:

Use the document corpus **idf** values in the query vector.

Example: query 1:

$q_{PC,j} = 0.30103$
$q_{games,j} = 0$
$q_{free,j} = 0$
$q_{download,j} = 0$

$$\sqrt{\sum_i q_{i,j}^2} = \sqrt{0.30103^2 + 0^2 + 0^2 + 0^2} = 0.30103$$

# Calculate Denominator (Alternative way)

$q_1$: PC games free download

$\sqrt{\sum_i w_{i,k}^2}$: Calculate as usual

| | D1 | D2 | D3 | D4 |
|---|---|---|---|---|
| **PC** | 0 | 0.30103 | 1.50515 | 0 |
| **More** | 0.30103 | 0 | 0.30103 | 0 |
| **Full** | 0 | 0.30103 | 0.60206 | 0 |
| **Shockwave** | 1.80618 | 0 | 0 | 0 |
| **Quality** | 0 | 0 | 0 | 0.60206 |

for $k = D1$ : $\sqrt{0.30103^2 + 1.80618^2}$
for $k = D2$ : $\sqrt{0.30103^2 + 0.30103^2}$
for $k = D3$ : $\sqrt{1.50515^2 + 0.30103^2 + 0.60206^2}$
for $k = D4$ : $\sqrt{0.60206^2}$

# Cosine Similarity (Alternative way)

## $q_1$: PC games free download

| | D1 | D2 | D3 | D4 |
|---|---|---|---|---|
| **PC** | 0 | 0.30103 | 1.50515 | 0 |
| **Games** | 0 | 0 | 0 | 0 |
| **Free** | 0 | 0 | 0 | 0 |
| **Download** | 0 | 0 | 0 | 0 |

$$sim(q_1, d_1) = \frac{0}{0.30103 \times \sqrt{0.30103^2 + 1.80618^2}} = \frac{0}{0.30103 \times 1.83109} = \mathbf{0}$$

$$sim(q_1, d_2) = \frac{0.30103}{0.30103 \times \sqrt{0.30103^2 + 0.30103^2}} = \frac{0.30103}{0.30103 \times 0.42572} = \mathbf{2.34896}$$

$$sim(q_1, d_3) = \frac{1.50515}{0.30103 \times \sqrt{1.50515^2 + 0.30103^2 + 0.60206^2}} = \frac{1.50515}{0.30103 \times 1.64881} = \mathbf{3.03249}$$

$$sim(q_1, d_4) = \frac{0}{0.30103 \times \sqrt{0.60206^2}} = \mathbf{0}$$

# Uninterpolated Average Precision

$$a_1 = \frac{\frac{1}{1}\frac{2}{2}\frac{3}{3}\frac{4}{4}}{4} = 1$$

$$a_2 = \frac{\frac{1}{1}\frac{2}{2}\frac{3}{3}\frac{4}{4}\frac{5}{5}\frac{5}{5}\frac{6}{6}\frac{7}{7}\frac{8}{8}\frac{9}{9}}{9} = 1$$

$$a_3 = \frac{\frac{1}{1}\frac{2}{2}\frac{3}{3}\frac{4}{4}\frac{5}{5}\frac{5}{5}\frac{6}{6}}{6} = 1$$

# Interpolated Average Precision

| Recall | Precision | Interp. Precision |
|---|---|---|
| 0.0 | 0.0 | |
| 0.1 | 1 | 1 |
| 0.2 | 1 | 1 |
| 0.3 | 0.75 | 0.875 |
| 0.4 | 0.8 | 0.875 |
| 0.5 | 0.83 | 0.875 |
| 0.6 | 0.857 | 0.875 |
| 0.7 | 0.875 | 0.875 |
| 0.8 | 0.727 | 0.75 |
| 0.9 | 0.75 | 0.75 |
| 1.0 | $0.\overline{6}$ | $0.\overline{6}$ |
| Avg Interp. Prec. | $\frac{9.541}{11} = 0.867$ | |