Web Information Retrieval

Dipl.-Inf. Christoph Carl Kling
Exercises WebIR

PLZ

ask questions!

WebIR@c-kling.de
Exercises WebIR

- No “homeworks”
- But exercises!
Proportabilities

A first stupid question (many more to come):

What is a “probability”? 
A first stupid question (many more to come):

What is a “probability”?

1) Frequentist: Tells us the expected frequency of events.

2) Bayesian: A personal belief on the outcome of events.
Exercise:

Prove the Bayes theorem using the definition of conditional probability.
Probabilistic retrieval:

Documents are binary vectors
Document collection

d1: Marcus tried to assassinate Caesar.
d2: Marcus was a Roman.
d3: Caesar was a ruler. All Romans were either loyal to Caesar or hated him.
d4: Everyone is loyal to someone. People only try to assassinate rulers they are not loyal to.
Document collection

d1: Marcus tried to assassinate Caesar.
d2: Marcus was a Roman.
d3: Caesar was a ruler. All Romans were either loyal to Caesar or hated him.
d4: Everyone is loyal to someone. People only try to assassinate rulers they are not loyal to.

Stopwords:

a, all, and, are, either, everyone, her, him, is, not, only, or, someone, they, to, was, were, who
Document collection

\[d1\]: Marcus tried assassinate Caesar .
\[d2\]: Marcus Roman .
\[d3\]: Caesar ruler. Romans loyal Caesar hated .
\[d4\]: loyal . People try assassinate rulers loyal .

Stopwords: 

\(a, all, and, are, either, everyone, her, him, is, not, only, or, someone, they, to, was, were, who\)
Document collection

d1: Marcus tried assassinate Caesar.
d2: Marcus Roman.
d3: Caesar ruler. Romans loyal Caesar hated.
d4: loyal. People try assassinate rulers loyal.

Stopwords:
a, all, and, are, either, everyone, her, him, is, not, only, or, someone, they, to, was, were, who

Stemming rules:
assassinate → assassin
assassinated → assassin
assassination → assassin
loyalty → loyal
hated → hate

Roman → Rome
Romans → Rome
ruler → rule
rulers → rule
tried → try
Document collection

d1: Marcus try assassin Caesar.
d2: Marcus Rome.
d3: Caesar rule. Rome loyal Caesar hate.
d4: loyal. People try assassin rule loyal.

Stopwords:

a, all, and, are, either, everyone, her, him, is, not, only, or, someone, they, to, was, were, who

Stemming rules:

assassinate → assassin
assassinated → assassin
assassination → assassin
loyality → loyal
hated → hate

Roman → Rome
Romans → Rome
ruler → rule
rulers → rule
tried → try
Document collection

d1: Marcus try assassin Caesar.
d2: Marcus Rome.
d3: Caesar rule. Rome loyal Caesar hate.
d4: loyal. People try assassin rule loyal.

Terms:
marcus, try, assassin, caesar, rome, rule, loyal, hate, people
Document collection

d1: Marcus try assassin Caesar .
d2: Marcus Rome .
d3: Caesar rule. Rome loyal Caesar hate .
d4: loyal . People try assassin rule loyal .

Terms:
marcus, try, assassin, caesar, rome, rule, loyal, hate, people

Binary vectors:

d1: (1, 1, 1, 1, 0, 0, 0, 0, 0, 0)
d2: (1, 0, 0, 0, 1, 0, 0, 0, 0, 0)
d3: (0, 0, 0, 1, 1, 1, 1, 1, 0, 0)
d4: (0, 1, 1, 0, 0, 1, 1, 0, 1)
Find the binary vector for the document:

\[ d5: \text{Someone assassinated Caesar.} \]

Stopwords:

- a, all, and, are, either, everyone, her, him, is, not, only, or, someone, they, to, was, were, who

Stemming rules:

- assassinate \rightarrow\text{assassin}
- assassinated \rightarrow\text{assassin}
- assassination \rightarrow\text{assassin}
- loyalty \rightarrow\text{loyal}
- hated \rightarrow\text{hate}
- Roman \rightarrow\text{Rome}
- Romans \rightarrow\text{Rome}
- ruler \rightarrow\text{rule}
- rulers \rightarrow\text{rule}
- tried \rightarrow\text{try}

Terms:

- marcus, try, assassin, caesar, rome, rule, loyal, hate, people
Probabilistic retrieval:

R: document is relevant

d: a document

What we need: p(R|d)
\[ \text{posterior} = \frac{\text{likelihood} \cdot \text{prior}}{\text{evidence}} \]
Probabilistic retrieval:

R: document is relevant
d: a document

What we need: \( p(R|d) \)

Human raters:

\( d_1: (1,1,1,1,0,0,0,0,0,0) \)
\( d_2: (1,0,0,0,1,0,0,0,0,0) \)
\( d_3: (0,0,0,1,1,1,1,1,0,0) \)
\( d_4: (0,1,1,0,0,1,1,0,1) \)
Probabilistic retrieval:

What we need: $p(R|d)$

Human raters:

$d_1$: $(1,1,1,1,0,0,0,0,0)$
$d_2$: $(1,0,0,0,1,0,0,0,0)$
$d_3$: $(0,0,0,1,1,1,1,1,0)$
$d_4$: $(0,1,1,0,0,1,1,0,1)$

Is $d_5$ relevant?
$d_5$: $(0,0,1,1,0,0,0,0,0)$
Probabilistic retrieval with term independence:

Assumptions

- Bag of words
- Independence of words
- Document relevance independent of other documents
Markov chain:

We are interested in ergodic Markov chains:

- homogeneous
- irreducible
- aperiodic
- positive recurrent
Markov Chain

We are interested in ergodic Markov chains:

- homogeneous (transition probabilities fixed)
- irreducible (every state always reachable)
- aperiodic (no greatest common divisor > 1 for recurrence)
We are interested in ergodic Markov chains:
We are interested in ergodic Markov chains:
We are interested in ergodic Markov chains:

![Diagram](image-url)
We are interested in ergodic Markov chains:
Entropy H

\[ A B C D E F G H I J K L M N \]
\[ \Xi \Omega \Theta \Xi I K L M N \]
\[ \Sigma T U \Phi \chi \psi \omega \]
\[\alpha \beta \gamma \delta \epsilon \zeta \eta \theta \iota \kappa \lambda \mu \nu \]
\[\xi \omicron \pi \rho \varsigma \sigma \tau \upsilon \phi \chi \psi \omega \]
Entropie & Co

Schmetterling und Taucherglocke

Ein Film von JULIAN SCHNABEL. Regisseur von «BATONI». «BEFORE NIGHT FALLS».

MATHIEU AMALRIC
EMMANUELLE SEIGNER
MARIJOSEE CROZE
ANN CONSIDINE

PATRICK CHRESSAI
NIELS ARENSTROEM
OLGA TATZ SARMAD
JEAN-PIERRE ZASSO
WALTER NOYER

www.film-ab.ch

Im Sonnenhof 3
8280 Kreuzlingen
Tel. 071 672 54 54

FESTIVAL DE CANNES
BESTE REGIE
JULIAN SCHNABEL
3 GOLDEN GLOBE NOMINIERUNGEN
4 OSCAR NOMINIERUNGEN

Entropie & Co

WeST C. C. Kling DM & ML 29 of 41
Distribution des lettres (%) dans un texte en français
<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.25</td>
</tr>
<tr>
<td>B</td>
<td>0.5</td>
</tr>
<tr>
<td>C</td>
<td>0.125</td>
</tr>
<tr>
<td>D</td>
<td>0.125</td>
</tr>
</tbody>
</table>
### Entropy & Co

<table>
<thead>
<tr>
<th></th>
<th>p</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0.125</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>0.125</td>
<td></td>
</tr>
</tbody>
</table>

\[
H(p) = - \sum_i p_i \cdot \log_2 p_i
\]
$H(p) = - \sum_{i} p_i \cdot \log_2 p_i$

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.25</td>
</tr>
<tr>
<td>B</td>
<td>0.5</td>
</tr>
<tr>
<td>C</td>
<td>0.125</td>
</tr>
<tr>
<td>D</td>
<td>0.125</td>
</tr>
</tbody>
</table>

$p($Character$)$ - bit per character
Entropy & Co

\[ H(p) = - \sum_{i} p_i \cdot \log_2 p_i \]

- 0.25
- 0.5
- 0.125
- 0.125

p(Charater) - bit per character

Entropy: 1.75
Entropy & Co

<table>
<thead>
<tr>
<th>p</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.25</td>
</tr>
<tr>
<td>B</td>
<td>0.5</td>
</tr>
<tr>
<td>C</td>
<td>0.125</td>
</tr>
<tr>
<td>D</td>
<td>0.125</td>
</tr>
</tbody>
</table>

Entropy: 1.75

Code:

- B
- A
- C
- D

Example text:
10110100
Entropy & Co

<table>
<thead>
<tr>
<th></th>
<th>P</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.25</td>
<td>0.125</td>
</tr>
<tr>
<td>B</td>
<td>0.5</td>
<td>0.25</td>
</tr>
<tr>
<td>C</td>
<td>0.125</td>
<td>0.125</td>
</tr>
<tr>
<td>D</td>
<td>0.125</td>
<td>0.5</td>
</tr>
</tbody>
</table>

\[
KL(P, Q) = \sum_{x \in X} P(x) \cdot (\log_2 P(x) - \log_2 Q(x))
\]

- \(p(\text{character in P})\)
- bit per character in P
- bit per character in Q
Thank you!
Questions?
Exercise:

Prove that cosine-similarity and Euclidean distance yield an identical relative distance measure (e.g. between documents) when all document and query vectors are normalised to be of length 1.
Probabilities

Independent events...