IN3120 week 11

Group 1

Reminder

- Aleksander ends next session with a Q&A. Ask questions!!
- I will go through the 2023 exam next week
  - Try to solve it yourself first

Tips for exam preparations|

- Reading the curriculum
  - Start with most common exam topics (check excel sheet)
  - Prioritize what's common in recent exams
  - Learn how to use the specific formulas!
- Do the exams
  - Set a 4h timer, try the exam without tools
  - Do an exam early, so you know what to focus on

Creds: gangen_i_tredje

Today’s format

- School exam topics occurring at least twice since 2018 (except TAAT)
- Each topic split into 2 sections

1. Theory/explanation
2. Last exam question when the topic was relevant

- Goal: explain the ideas without solving the exam questions
- Look out for «*the italic text*» (explains the intuition)

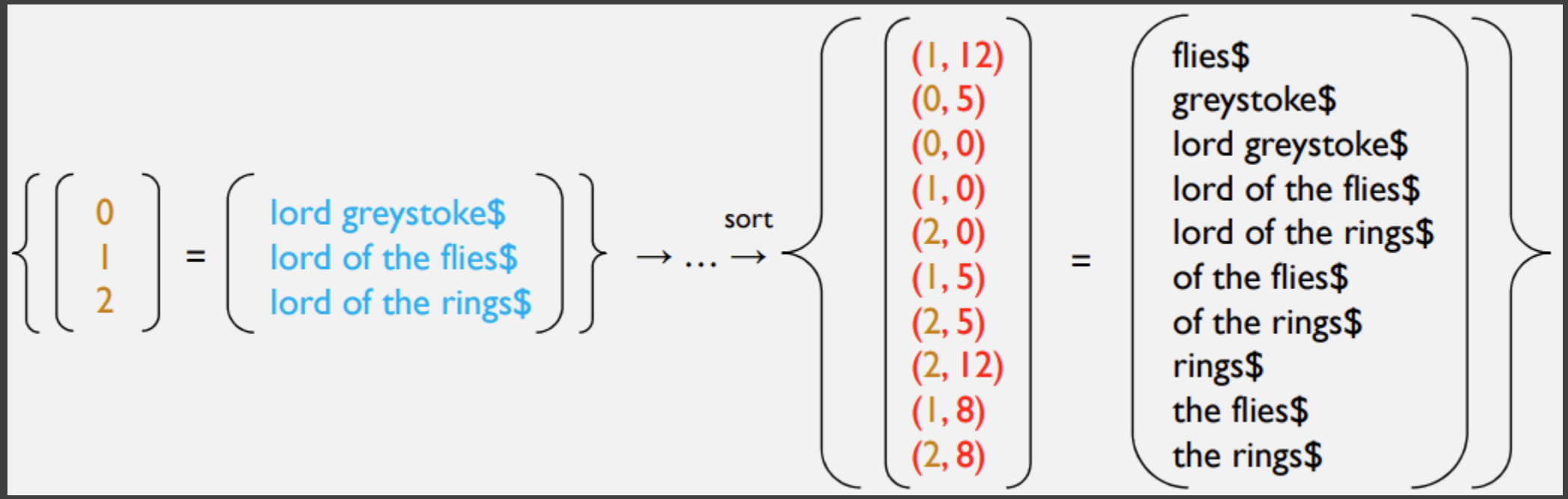
Agenda

- Recap of common exam topics
  - Suffix arrays
  - SVMs
  - MAP
  - Cosine similarity
  - Precision & recall, F1-score
  - Random surfer model
  - (~~Term-at-a-time~~)
  - Bloom filter
  - Kendall Tau
- Shoutout of the week

Suffix array

1. Split the buffers on each term/character
    1. The positions where we split are the offsets
2. Store as (docId, offset)-pairs
3. Sort lexicographically

Suffix array



Why bother?

1. We can perform binary search (low complexity)
2. We don’t need to store the text buffers more than once, only the offset positions in the buffers
   - When the array is sorted and we’re performing binary search, we don’t care about the ‘other half’, since it won’t match
   - *«If I know it’s a mismatch, I don’t need to read it»*

Suffix arrays – exam 2022

2 SUFFIX ARRAYS [15p]

Consider the questions and statements below about suffix arrays, and select the correct answer. Justify your answers, i.e., explain why you believe that your answer is correct. Answers with no justification give no credit.

a) [5p] What is the suffix array for the string *engineering*?

1. [2, 3, 8, 4, 9, 1, 7, 5, 0, 6, 10]
2. [5, 0, 10, 6, 2, 3, 8, 4, 9, 1, 7]
3. [5, 0, 6, 10, 2, 4, 9, 1, 7, 3, 8]
4. [5, 0, 6, 10, 2, 3, 8, 4, 9, 1, 7]
5. [5, 10, 0, 6, 8, 3, 2, 4, 9, 7, 1]

b) [5p] In IN3120's obligatory assignment on suffix arrays, the naïve comparison-based sorting algorithm used to construct the suffix array for a string s runs in $O(n\log n)$ time, where $|s| = n$.

1. True
2. False

c) [5p] Using the suffix array for a string s with $|s| = n$, what is the best bound required to locate in s the first occurrence of a pattern having length m, where $m < n$?

1. $O(nm)$
2. $O(n^2)$
3. $O(mn\log n)$
4. $O(m\log n)$
5. $O(\log n)$

Agenda

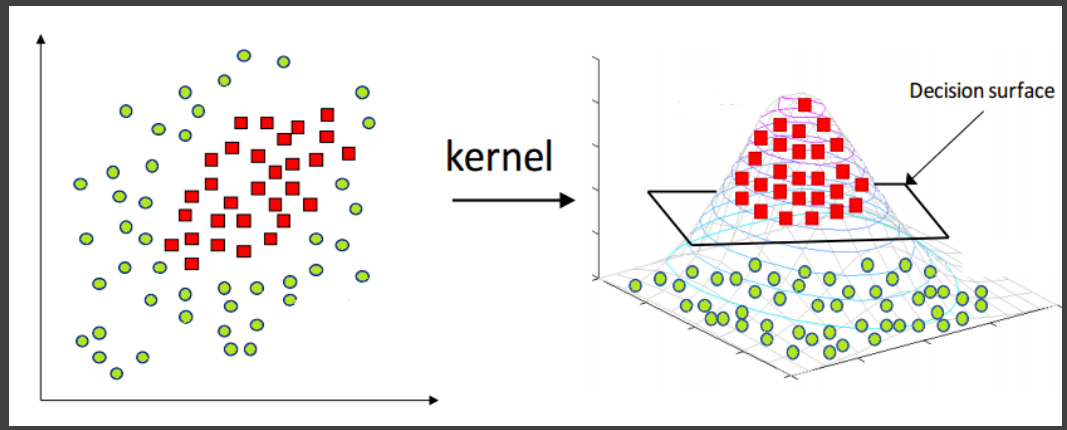
- Recap of common exam topics
  - Suffix arrays
  - SVMs
  - MAP
  - Cosine similarity
  - Precision & recall, F1-score
  - Random surfer model
  - Bloom filter
  - Kendall Tau
- Shoutout of the week

Support vector machines

- We need a decision boundary for classification
- Idea: Create a hyperplane that maximizes the distance between the classes
- Support vectors: The data points closest to the decision boundary
  - *«The support vectors are the nodes closest to the other class»*

The Kernel trick

- If the data is not linearly separable, use the Kernel trick
- Raise the nodes into a higher dimension until the classes are linearly separable



Support vector machines – exam 2022

e) [4p] In SVMs the concept of a support vector is key. Describe what a support vector is, and which role support vectors play.

Agenda

- Recap of common exam topics
  - Suffix arrays
  - SVMs
  - MAP
  - Cosine similarity
  - Precision & recall, F1-score
  - Random surfer model
  - Bloom filter
  - Kendall Tau
- Shoutout of the week

Mean Average Precision

- MAP: Average average precision
- We want to measure search engine quality
- A measure of how well the engine ranks relevant documents on average
- *«How well-placed are my relevant documents in an average query?»*

MAP formula

1. Perform Precision@k for all relevant docs
2. Average of (1)
3. Average of (2) for for multiple queries

Example

- Given 4 query results
- q1: [R, N, R, N, N, R, R]
- q2: [N, R, R, R, N, N, N]
- q3: [R, R, R, R, N, R, N]
- q4: [R, N, R, N, R, N, R]

Example – precision@k (1)

- Given 4 query results
- q1: [R, N, R, N, N, R, R] → [1/1, N, 2/3, N, N, 3/6, 4/7]
- q2: [N, R, R, R, N, N, N] → [N, 1/2, 2/3, 3/4, N, N, N]
- q3: [R, R, R, R, N, R, N] → [1/1, 2/2, 3/3, 4/4, N, 5/6, N]
- q4: [R, N, R, N, R, N, R] → [1/1, N, 2/3, N, 3/5, N, 4/7]

Example – Average precision (2)

- Given 4 query results
- q1: [R, N, R, N, N, R, R] → 2.738 / 4 → 0.6845
- q2: [N, R, R, R, N, N, N] → 1.917 / 3 → 0.639
- q3: [R, R, R, R, N, R, N] → 4.833 / 5 → 0.9666
- q4: [R, N, R, N, R, N, R] → 2.838 / 4 → 0.7095

Example – Mean average precision (3)

- Given 4 query results
- q1: [R, N, R, N, N, R, R] → 2.738 / 4 → 0.6845
- q2: [N, R, R, R, N, N, N] → 1.917 / 3 → 0.639
- q3: [R, R, R, R, N, R, N] → 4.833 / 5 → 0.9666
- q4: [R, N, R, N, R, N, R] → 2.838 / 4 → 0.7095
- MAP → (0.6845 + 0.639 + 0.9666 + 0.7095) / 4 = **0.7499**

MAP – exam 2023

2 MEASURING RELEVANCE [20p]

(a) [5p] Describe what the F_β-score is, and define it in terms of precision P and recall R. What does the β parameter control? If $P = 0.1$ and $R = 0.5$, what is the F_1-score?

(b) [5p] Assume a ranked retrieval context. Describe what a precision-recall curve is and how we generate it. What is an interpolated precision-recall curve?

(c) [5p] Let R denote a relevant document, and let N denote a non-relevant document. Consider a search system that for the query *carrot* produces the ranked result set RRNRNNNR and that for the query *chocolate* produces the ranked result set RNRR. Show how you compute the search system's mean average precision (MAP) score. (Clearly showing the correct steps without arriving at a final numerical result will give full marks.)

(d) [5p] Kendall's tau distance can help us assess how "close" a ranked result set L for a given query is to a given set of pairwise preferences P for that query. Describe the high-level idea behind how this is computed. If L = [A, C, B, D] and P = {(A, B), (A, C), (A, D), (B, C), (B, D), (C, D)}, what is Kendall's tau distance between the two?

Agenda

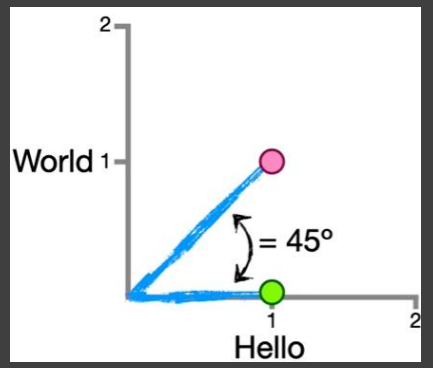
- Recap of common exam topics
  - Suffix arrays
  - SVMs
  - MAP
  - Cosine similarity
  - Precision & recall, F1-score
  - Random surfer model
  - Bloom filter
  - Kendall Tau
- Shoutout of the week

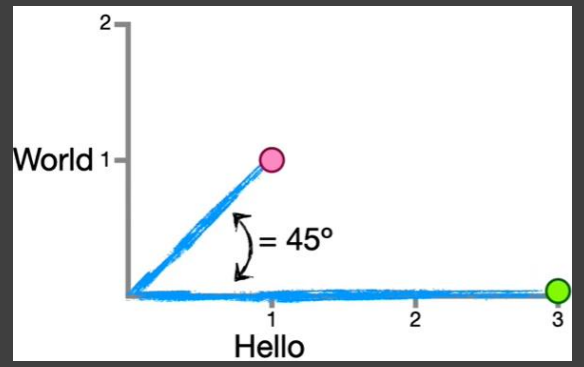
Cosine similarity

- Used for measuring similarity between text in a vector space
- Why not measure by distance?
  - Repeating terms can make it look like documents are less similar
- Use cosine of angles instead:
  - If cosine similarity is 1, documents are identical
  - If cosine similarity is 0, documents have nothing in common

Cosine similarity vs. euclidean distance

- Let’s say we have the documents
  - Hello World
  - Hello
  - Hello Hello Hello
- Large distance, but the same degree between vectors





- Check out StatQuest: https://www.youtube.com/watch?v=e9U0QAFbfLI&ab_channel=StatQuestwithJoshStarmer

Cosine similarity – exam 2023

1 VECTOR SPACES [35p]

(a) [10p] There are two main ways to think about vector spaces for text: (A) As a sparse and extremely high-dimensional representation where each unique vocabulary term corresponds to a distinct dimension of your vector space, or (B) as a dense and lower-dimensional representation where each unique word in your vocabulary corresponds to a point in an abstract vector space we can call an "embedding space" (where we beforehand have fixed, say, typically a few hundred dimensions.)

- (i) For case (A) above, briefly discuss how a larger text buffer (e.g., a document) could be placed in this vector space, and outline some pros and cons of working with representation (A).
- (ii) For case (B) above, briefly and at a high level discuss the general ideas behind how a given word gets placed in this embedding space, how we might go about placing a larger text buffer (e.g., a document) in this same embedding space, and outline some pros and cons of working with representation (B).

(b) [5p] Consider the dense vectors x = [0.6, 0.2, 0.8] and y = [1.0, 0.1, 0.9]. Show how to compute the cosine similarity between x and y. (Clearly showing the correct procedure without arriving at a final numerical result will give full marks.)

(c) [5p] Explain what an approximate nearest neighbour (ANN) index is, and why it is useful.

(d) [15p] List at least 5 strategies that an ANN index can employ to efficiently find matches, and succinctly explain the thinking behind each strategy.

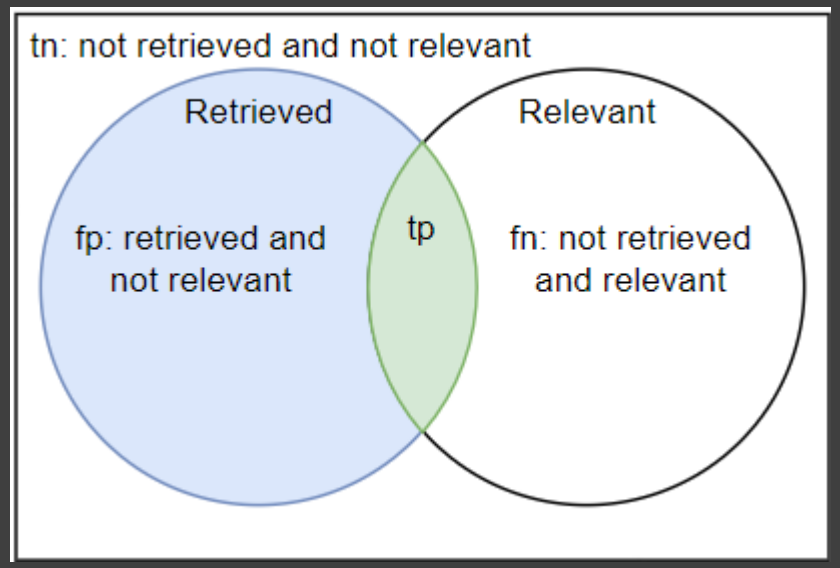
Agenda

- Recap of common exam topics
  - Suffix arrays
  - SVMs
  - MAP
  - Cosine similarity
  - Precision & recall, F1-score
  - Random surfer model
  - Bloom filter
  - Kendall Tau
- Shoutout of the week

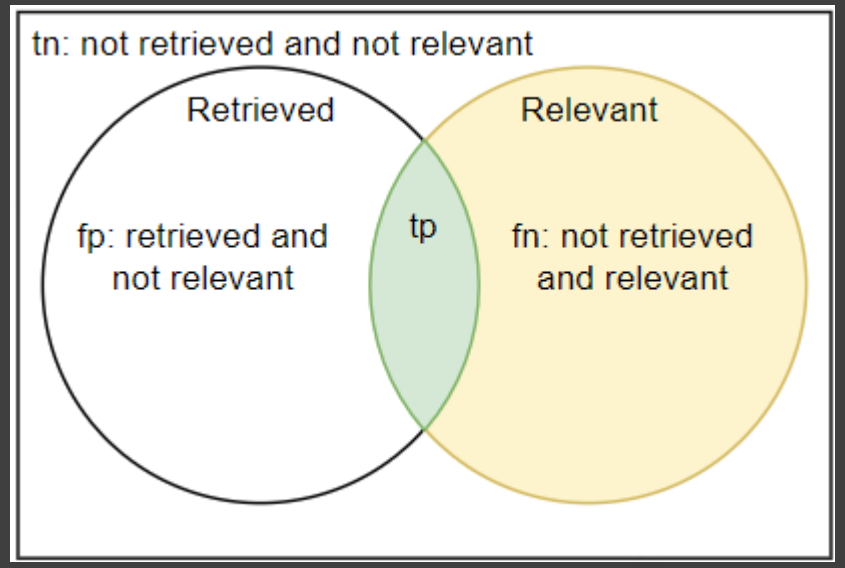
Precision & Recall

- Precision: *«Of all retrieved documents, how many are relevant?»*
- Recall: *«Of all relevant documents, how many are retrieved?»*

$$P = \frac{tp}{tp + fp} = P(relevant|retrieved)$$



$$R = \frac{tp}{tp + fn} = P(retrieved|relevant)$$



More Precision & Recall…

- Precision: *«If i pick a random retrieved document, what is the probability is is relevant»*

$$P = P(relevant|retrieved)$$

- Recall: *«If i pick a random relevant document, what is the probability I retrieved it?»*

$$R = P(retrieved|relevant)$$

F-score

- Weighted harmonic mean of precision and recall

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R} \quad \text{where} \quad \beta^2 = \frac{1-\alpha}{\alpha}$$

- Use β to edit the priority of precision and recall
  - β < 1emphasizes precision, β > 1 emphasizes recall

- section 8.5, 8.6: https://nlp.stanford.edu/IR-book/pdf/08eval.pdf

F-score edge-case example

β < 1emphasizes precision, β > 1 emphasizes recall

$$F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

$$P = 0.10, R = 0.90$$

$$F_{\beta=1000} = \frac{(1000^2 + 1) * 0.10 * 0.90}{1000^2 * 0.10 + 0.90} = \frac{90\,000.09}{100\,000.9} = 0.8999$$

$$F_{\beta=0.001} = \frac{(0.001^2 + 1) * 0.10 * 0.90}{0.001^2 * 0.10 + 0.90} = \frac{0.09000009}{0.9000001} = 0.10000008$$

F_1-score

- Tip: learn F_1-score before F-score
  - But you should learn both!
- F_1-Score is when *p* and *r* have equal emphasis

$$F = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \; where \; \beta^2 = \frac{1 - \alpha}{\alpha}$$

$$if \; \alpha = 0.5 \; then \; \beta = 1$$

$$F_{\beta=1} = \frac{(1^2 + 1)PR}{1^2 P + R} = \frac{2PR}{P + R}$$

Confusion matrix

	Actual		
Predicted		Positive	Negative
	Positive	True positive (tp)	False positive (fn)
	Negative	False negative (fn)	True negative (tn)

Precision, recall & F1-score example

- We have some values. What is the Precision, Recall and F_1-Score?
- Let's crunch come numbers!

	Actual		
Predicted		Positive	Negative
	Positive	210	30
	Negative	10	110

Precision, recall & F1-score example

$$P = \frac{tp}{tp + fp} = \frac{210}{210 + 30} = 0.875$$

$$R = \frac{tp}{tp + fn} = \frac{210}{210 + 10} = 0.955$$

$$F_1 = \frac{2PR}{P + R} = \frac{2 * 0.875 * 0.955}{0.875 + 0.955} = \frac{1.671}{1.83} = 0.913$$

	Actual		
Predicted		Positive	Negative
	Positive	210	30
	Negative	10	110

Precision & Recall, F1 – exam 2023

2 MEASURING RELEVANCE [20p]

(a) [5p] Describe what the F_β-score is, and define it in terms of precision P and recall R. What does the β parameter control? If $P = 0.1$ and $R = 0.5$, what is the F_1-score?

(b) [5p] Assume a ranked retrieval context. Describe what a precision-recall curve is and how we generate it. What is an interpolated precision-recall curve?

(c) [5p] Let R denote a relevant document, and let N denote a non-relevant document. Consider a search system that for the query *carrot* produces the ranked result set RRNRNNNR and that for the query *chocolate* produces the ranked result set RNRR. Show how you compute the search system's mean average precision (MAP) score. (Clearly showing the correct steps without arriving at a final numerical result will give full marks.)

(d) [5p] Kendall's tau distance can help us assess how "close" a ranked result set L for a given query is to a given set of pairwise preferences P for that query. Describe the high-level idea behind how this is computed. If L = [A, C, B, D] and P = {(A, B), (A, C), (A, D), (B, C), (B, D), (C, D)}, what is Kendall's tau distance between the two?

Agenda

- Recap of common exam topics
  - Suffix arrays
  - SVMs
  - MAP
  - Cosine similarity
  - Precision & recall, F1-score
  - Random surfer model
  - Bloom filter
  - Kendall Tau
- Shoutout of the week

Random Surfer model

- Our imaginary web surfer will either
  - Follow a direct link, or
  - Teleport randomly

- Used for calculating PageRank
  - The fraction of time spent on a website after infinite surfing

Surfing probabilities

- Let's define some probabilities using a teleportation rate α
  - Following any direct link: $(1-\alpha)$
  - Performing any teleportation: α
  - Following a specific link: $\frac{1-\alpha}{|direct\ links|}$
  - Performing specific teleportation: $\frac{\alpha}{|websites\ not\ linked|}$
  - We will always either teleport or follow a direct link

$$P(teleport) + P(direct\ link) = \alpha + (1-\alpha) = 1$$

  - $|direct\ links|$ means the number of direct links

Let’s go surfing

- Let’s say $\alpha = 0.1$ (10% chance of teleporting randomly)
- We have 5 direct links and 10 pages not linked

- Following any direct link: $(1 - 0.1) = 0.9$
- Performing any teleportation: 0.1

- Following a specific link: $\frac{1-0.1}{5} = 0.18$
- Performing specific teleportation: $\frac{0.1}{10} = 0.01$
- Summing the probabilites: $5 * 0.18 + 10 * 0.01 = 1$

Random Surfer model – exam 2022

c) [4p] In the random surfer model the surfer is allowed to teleport with some probability greater than zero. If you assume that teleportation is not allowed, discuss some of the problems that can arise and their implications.

Agenda

- Recap of common exam topics
  - Suffix arrays
  - SVMs
  - MAP
  - Cosine similarity
  - Precision & recall, F1-score
  - Random surfer model
  - Bloom filter
  - Kendall Tau
- Shoutout of the week

Bloom filter

- Probabilistic way of searching for set membership
- Can never guarantee that the set contains an element, but we can guarantee if it doesn’t
- Uses k hash functions to load a bit vector with 1’s

Building the filter

- Let’s say we have
  - 2 hash functions, H_1 and H_2
  - The set {informatikk, er, gøy}

Building the filter

index	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
bit-vector	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Keep in mind

- In practice, it doesn’t make sense to have a 20-bit vector and 3 elements in the set
- A bloom-filter is only useful if we have more terms than bits. Otherwise, we could assign 1 bit for each term

Building the filter

- H_1(Informatikk) → 10
- H_2(Informatikk) → 4

Building the filter

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0

Building the filter

- H_1(informatikk) → 10
- H_2(Informatikk) → 4

- H_1(er) → 6
- H_2(er) → 7

Building the filter

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
0	0	0	0	1	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0

Building the filter

- H_1(informatikk) → 10
- H_2(Informatikk) → 4

- H_1(er) → 6
- H_2(er) → 7

- H_1(gøy) → 6
- H_2(gøy) → 13

Building the filter

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
0	0	0	0	1	0	1	1	0	0	1	0	0	1	0	0	0	0	0	0

Testing the filter

- H_1(Aleksander) → 3
- H_2(Aleksander) → 13
- Returns: Guaranteed not in the set

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
0	0	0	0	1	0	1	1	0	0	1	0	0	1	0	0	0	0	0	0

Testing the filter

- H_1(Truls) → 6
- H_2(Truls) → 10
- Returns: Probably in the set
  - This is wrong! That's why it's probabilistic (probably correct)

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
0	0	0	0	1	0	1	1	0	0	1	0	0	1	0	0	0	0	0	0

Bloom filter – exam 2023

(d) [5p] Consider a tiny Bloom filter backed by 16 bits of storage and with 3 hash functions. Assume the hash values shown in the table below.

- (i) Show what the filter's bit array looks like before inserting anything, after inserting *carrot*, and after inserting both *carrot* and *toffee*.
- (ii) Given that only the two values *carrot* and *toffee* have been inserted, explain what the filter will say when queried about the set memberships of *steak* and *carrot*, respectively, and explain the logic for how the filter arrives at these decisions.
- (iii) Outline how you could modify the Bloom filter to reduce the probability of false positives.

Hash function *h*	**Value *x***	**Hash value *h*(*x*)**
h_1	*carrot*	12
h_1	*toffee*	0
h_1	*steak*	7
h_2	*carrot*	7
h_2	*toffee*	12
h_2	*steak*	15
h_3	*carrot*	15
h_3	*toffee*	3
h_3	*steak*	11

Agenda

- Recap of common exam topics
  - Suffix arrays
  - SVMs
  - MAP
  - Cosine similarity
  - Precision & recall, F1-score
  - Random surfer model
  - Bloom filter
  - Kendall Tau
- Shoutout of the week

Kendall Tau distance

- Measures the quality of pairwise preference in our search engine
- *«A document is either more or less relevant than any other document. How much does our search engine agree with this ranking?»*

$$KT = \frac{X - Y}{X + Y}$$

Example (same as week 7)

- Pairwise preference: {(1, 2), (1, 3), (1, 4), (2, 3), (2, 4), (3, 4)}
- 1 > 2 2 > 3 3 > 4
- 1 > 3 2 > 4
- 1 > 4

Example (same as week 7)

- X: Agreements
- Y: Disagreements

$$KT = \frac{X - Y}{X + Y}$$

KT Distance examples

- [1, 3, 2, 4]: (5 – 1) / (5 + 1) = 0,67
- [4, 3, 2, 1]: (0 – 6) / (0 + 6) = -1.0
- [1, 4, 3, 2]: (3 – 3) / (3 + 3) = 0

Kendall Tau distance – exam 2023

2 MEASURING RELEVANCE [20p]

(a) [5p] Describe what the F_β-score is, and define it in terms of precision P and recall R. What does the β parameter control? If $P = 0.1$ and $R = 0.5$, what is the F_1-score?

(b) [5p] Assume a ranked retrieval context. Describe what a precision-recall curve is and how we generate it. What is an interpolated precision-recall curve?

(c) [5p] Let R denote a relevant document, and let N denote a non-relevant document. Consider a search system that for the query *carrot* produces the ranked result set RRNRNNNR and that for the query *chocolate* produces the ranked result set RNRR. Show how you compute the search system's mean average precision (MAP) score. (Clearly showing the correct steps without arriving at a final numerical result will give full marks.)

(d) [5p] Kendall's tau distance can help us assess how "close" a ranked result set L for a given query is to a given set of pairwise preferences P for that query. Describe the high-level idea behind how this is computed. If L = [A, C, B, D] and P = {(A, B), (A, C), (A, D), (B, C), (B, D), (C, D)}, what is Kendall's tau distance between the two?

Agenda

- Recap of common exam topics
  - Suffix arrays
  - SVMs
  - MAP
  - Cosine similarity
  - Precision & recall, F1-score
  - Random surfer model
  - Bloom filter
  - Kendall Tau
- Shoutout of the week

Shoutout: Max Martin

- Swedish songwriter
- Makes the worlds most famous pop songs
- 2nd most number-one singles ever
- Has 30 songs on Spotify with > 1.000.000.000 streams

27 Billboard top 100 singles

1. 1998 – "...Baby One More Time" by Britney Spears
2. 2000 – "It's Gonna Be Me" by NSYNC
3. 2008 – "I Kissed a Girl" by Katy Perry
4. 2008 – "So What" by Pink
5. 2009 – "My Life Would Suck Without You" by Kelly Clarkson
6. 2009 – "3" by Britney Spears
7. 2010 – "California Gurls" by Katy Perry featuring Snoop Dogg
8. 2010 – "Teenage Dream" by Katy Perry
9. 2010 – "Raise Your Glass" by Pink
10. 2011 – "Hold It Against Me" by Britney Spears
11. 2011 – "E.T." by Katy Perry featuring Kanye West
12. 2011 – "Last Friday Night (T.G.I.F.)" by Katy Perry
13. 2012 – "Part of Me" by Katy Perry
14. 2012 – "One More Night" by Maroon 5
15. 2012 – "We Are Never Ever Getting Back Together" by Taylor Swift
16. 2013 – "Roar" by Katy Perry
17. 2013 – "Dark Horse" by Katy Perry featuring Juicy J
18. 2014 – "Shake It Off" by Taylor Swift
19. 2014 – "Blank Space" by Taylor Swift
20. 2015 – "Bad Blood" by Taylor Swift featuring Kendrick Lamar
21. 2015 – "Can't Feel My Face" by The Weeknd
22. 2016 – "Can't Stop the Feeling!" by Justin Timberlake
23. 2019 – "Blinding Lights" by The Weeknd
24. 2021 – "Save Your Tears" by The Weeknd and Ariana Grande
25. 2021 – "My Universe" by Coldplay and BTS
26. 2024 – "Yes, And?" by Ariana Grande
27. 2024 – "We Can't Be Friends (Wait for Your Love)" by Ariana Grande

15 min break