Assignment 05

Digital Libraries and Foundations of Information Retrieval

Winter semester 2022

1542011 Franka Brunen, 1365848 Andreas Schneider

Task 1: Thesauri 2+5+2+3+3 Points

(a)
1 S: (n) queen (the only fertile female in a colony of social insects such as bees and ants and termites; its function is to lay eggs)
2 S: (n) queen, queen regnant, female monarch (a female sovereign ruler)
3 S: (n) queen (the wife or widow of a king)
4 S: (n) queen (something personified as a woman who is considered the best or most important of her kind)
5 „Paris is the queen of cities"; „the queen of ocean liners"
6 S: (n) king, queen, world-beater (a competitor who holds a preeminent position)
7 S: (n) fagot, faggot, fag, fairy, nance, pansy, queen, queer, poof, poove, pouf (offensive term for a homosexual man)
8 S: (n) queen (one of four face cards in a deck bearing a picture of a queen)
9 S: (n) queen ((chess) the most powerful piece)
10 S: (n) queen, queen mole rat (an especially large mole rat and the only member of a colony of naked mole rats to bear offspring which are sired by only a few males)
11 S: (n) tabby, queen (female cat)

(b)
(1) Meaning 8
(2) Meaning 2
(3) Meaning 4, 5, 6
(4) Meaning 1
(5) Meaning 9
(6) None

Sometimes a thesaurus has more than one distinct meaning to a word, like in sentence (3). Also entities, like „The queen" (of England) in sentence (2) and „Queen" (band) in sentence (6) are not covered in thesauri.

(c) Meanings 2 and 3 are handled in *Monarchy*, 8 and 9 in *Gaming*, 1 and 11 in *Science*, and 7 in *Other uses*.

(d) That approach would always assign the same meaning to all occurrences of a word. Many words have very different meanings depending on the context. The most frequent meaning also has a high probability of being one of the more generic ones. Even if „run" was identified as a verb, the approach would mismatch either „The company is running fine", or „The company has not been run down".

(e) The assignment could be improved by involving context. Just like a „John Doe" is only recognizable as a distinct author when including co-authors, locations, publishers, time frames etc. in a disambiguation, the assignment of term occurrences to thesaurus entries will become less ambiguous when including semantic context. For example if „Queen" is referred to as „she", the music band is already ruled out.

Task 2: String Distances 2+4+9 Points

(a) The Hamming Distance is the amount of characters which would need to be replaced to transform one string into the other. Both strings need to have the same length.[1]

(b) The Levenshtein Distance is a measure of how many single-character edit operations are required to transform one string into the other. Operations are *insert*, *delete* and *substitute*. The variant by Damerau adds the operation of *transposition*, the exchange of two neighboring characters, at the same cost as the other operations in its regular form.[2]

(c) The Levenshtein Distances in this exercise are unweighted.

rhein and **rhine**

Hamming Distance	rhein		0
	rh**i**in	replace **e** with **i**	1
	rhe**n**n	replace **i** with **n**	2
	rhin**e**	replace **n** with **e**	**3**
Levenshtein Distance	rhein		0
	rhein	delete **e**	1
	rh in**e**	insert **e**	**2**
Damerau-Levenshtein Distance	no difference		

mainz and **minze**

Hamming Distance	mainz		0
	m**i**inz	replace **a** with **i**	1
	mi**n**nz	replace **i** with **n**	2
	min**z**z	replace **n** with **z**	3
	minz**e**	replace **z** with **e**	**4**
Levenshtein Distance	mainz		0
	m~~a~~inz	delete **a**	1
	m inz**e**	insert **e**	**2**
Damerau-Levenshtein Distance	no difference		

abababab and **babababa**

Hamming Distance	abababab		0
	bbababab	replace **a** with **b**	1
	b**a**ababab	replace **b** with **a**	2
	...		
	bababab**a**	replace **b** with **a**	**8**
Levenshtein Distance	abababab		0
	~~a~~bababab	delete **a**	1
	bababab**a**	insert **a**	1
Damerau-Levenshtein Distance	no difference		

[1] https://cgl.ethz.ch/teaching/former/infotheory0607/Downloads/hamming1_1.pdf

[2] https://suw.biblos.pk.edu.pl/resources/i5/i0/i8/i0/i3/r50803/NiewiarowskiA_ParallelizationLevenshtein.pdf

Task 3: Context-sensitive Spelling Correction 2+8+5 Points

(a) $P(miles|thousand), P(males|thousand), P(mules|thousand)$

(b) Using Firefox, private window, google.com. The variable C (C is the number of distinct bigrams in the corpus with w_i at the first position) is left out for the purpose of simplicity.

About 2.610.000.000 results for `thousand`.

About 13.300.000 results for „`thousand miles`“. $P(miles|thousand) = \frac{13.3\text{E}6}{26.1\text{E}8} = 0.51\%$

About 61.800 results for „`thousand males`“. $P(males|thousand) = \frac{61800}{26.1\text{E}8} = 0.0024\%$

About 9.200 results for „`thousand mules`“. $P(mules|thousand) = \frac{9200}{26.1\text{E}8} = 0.00035\%$

`miles` is the best correction term out of the *three* candidates.

(c) Possible information need: „Does the ZIMK host an event for christmas?“

`zimk`, as an abbreviation, could be misinterpreted as a human spelling error.

Both duckduckgo.com and google.com included results for `zink` and `zinc`. Google announced it („Did you mean *zink* christmas?“), Duckduckgo did not.

I am pretty sure that more people will have mistyped `zink/zinc` in this query „`zimk christmas`“, looking for decoration, cups etc., than people who actually want to know about the christmas plans of whatever `zimk` means to them. That means that more than half of the time, people querying for „`zimk christmas`“ are helped by including search results with the most probable correction terms.