Digital Libraries WS 2018/2019
Übungsblatt 4

Aaron Winziers - 1176638; Michael Wolz - 1195270

30. November 2018

Aufgabe 1

a)

1. **water ice** (water frozen in the solid state
2. (the frozen part of a body of water)
3. **sparkler** (diamonds)
4. **frosting, icing** (a flavored sugar topping used to coat and decorate cakes)
5. **methamphetamine, methamphetamine hydrochloride, Methedrine, meth, deoxyephedrine, chalk, chicken feed, crank, glass, shabu, trash (an amphetamine derivative** (trade name Methedrine) used in the form of a crystalline hydrochloride; used as a stimulant to the nervous system and as an appetite suppressant)
6. **internal-combustion engine, ICE** (a heat engine in which combustion occurs inside the engine rather than in a separate furnace; heat expands a gas that either moves a piston or turns a gas turbine)
7. **ice rink, ice-skating rink** (a rink with a floor of ice for ice hockey or ice skating) "the crowd applauded when she skated out onto the ice"

b)

1. 1, 2
2. 1
3. 6

4. Keiner der Bedeutungen
5. 1, 4

Ich verstehe die Aufgabenstellungen einfach nicht und überlasse den Rest dir ;*

Aufgabe 2

a)

Die Hamming-Distanz auch Hamming-Abstand oder -Gewicht misst wie Unterschiedlich zwei zu vergleichende Strings sind. Die Distanz zwischen zwei Strings der gleichen Länge ist die Anzahl der Zeichen die ersetzt werden müssen in einem String um das andere zu erzeugen bzw. wie viele Zeichen unterschiedlich sind.

Quelle: https://sciencing.com/how-to-calculate-hamming-distance-12751770.html

b)

Levenshtein-Distanz ist ein Vergleich zwischen zwei Strings der angibt wie viele Operationen auf einzelne Buchstaben benötigt werden um ein String in das andere umzuwandeln. Hierbei sind die Operationen: Einfügen, Löschen und Ersetzen.

Damerau-Levenshtein-Distanz ist im Prinzip das gleiche Algorithmus wie die Levenshtein-Distanz nur mit einer Ergänzung durch einen Tausch-Operator der zwei vertauschte Zeichen wechseln kann (Strign → String)

Quelle: https://www.cs.helsinki.fi/u/tpkarkka/opetus/13s/spa/lecture07.pdf

c)

lewenstein → levenshtein

- Hamming-Distanz = 6
- Levenshtein-Distanz = 2
- Damerau-Levenshtein-Distanz = 2

trier → tire

- Hamming-Distanz = 3
- Levenshtein-Distanz = 3

- Damerau-Levenshtein-Distanz = 2

10101010 → 01010101

- Hamming-Distanz = 8
- Levenshtein-Distanz = 2
- Damerau-Levenshtein-Distanz = 2

d)

- Schulz = S420
- Scholz = S420
- Schaeuble = S140
- Chewbacca = C120
- Chewie = C000

Laut dem SOUNDEX Algorithmus werden Schulz und Scholz auf Englisch gleich ausgesprochen. Chewie und Chewbacca sind im Vergleich nicht ganz so naheliegend, jedoch sind sie immer noch zu einender mehr ähnlich als Schaeuble zu allen anderen Namen.

Aufgabe 3

a)

Benötigte bedingte Wahrscheinlichkeiten:

- $P(gravity|blame)$
- $P(brevity|blame)$
- $P(for|gravity)$
- $P(for|brevity)$

b)

- $P(gravity|blame) = \frac{18.900.000}{315.000.000} = 0,06$
- $P(brevity|blame) = \frac{1.610.000}{315.000.000} = 0,005$
- $P(for|gravity) = \frac{375.000.000}{1.440.000.000} = 0,26$
- $P(for|brevity) = \frac{12.500.000}{12.500.000} = 1$

$P(\text{blame gravity for}) = 0,0156$

$P(\text{blame gravity for}) = 0,005$

Das heißt dass "blame gravtiy for"das bessere der Beiden Kandidaten ist. Die Wahrscheinlichkeiten wurden mit Google berechnet obwohl ein anderer Korpus besser wäre weil an Quellen so wie der Corpus of Contemporary American English und wordandphrase.info keine Ergebnisse für die suchen "blame gravityünd "blame brevity"kamen.

c)

Ein Student oder Mitarbeiter der Uni Trier könnte wissen wollen ob das Zentrum für Informations-, Medien- und Kommunikationstechnologie der Universität über den Weihnachtsferien offen ist und zur Verfügung steht.

Hier könnte als Problem bei einer automatischen Rechtschreibekorrektur vorkommen dass eine Suchmaschine zimk zu Zimt korrigiert und Rezepte für weihnachtliches Zimtgebäck o.Ä. aufrufen obwohl diese nicht dem ursprünglichen Information Need entsprechen.

Bei Google wird zimk zu Zinc automatisch korrigiert. Dieser Vorschlag kommt möglicherweise vor weil Zinc und Zimk phonetisch serh ähnlich sind oder auch weil auf den meisten Tastaturen das m und das n sehr nah aneinander liegen.