

Tentamen Programmeren voor taal en tekst

20 juni 2011

Geef duidelijke antwoorden, maar hou het kort en bondig. Er zijn 10 vragen. Elke vraag is 1 punt waard.

1. (a) Leg uit wat wordt bedoeld met de type/token-ratio van een tekst.
(b) Hoe verhoudt zich de type/token-ratio van een korte tekst zich meestal tot de type/token-ratio van een lange tekst? Is een van de twee meestal groter dan de ander of zijn ze meestal even groot? Verklaar je antwoord.
2. (a) Leg uit wat lemmatisering is.
(b) Is het mogelijk dat de frequentie van een bepaald woord in een corpus anders is dan de frequentie van een lemma met exact dezelfde vorm in hetzelfde corpus? Verklaar je antwoord.
3. (a) Laat zien hoeveel woord-3-grammen een zin met zes woorden bevat, bijvoorbeeld voor: *dit is een hele goede vraag*
(b) Beschrijf kort een automatische methode om alle woord-3-grammen in een tekst op te sporen.
4. (a) Noem 3 methoden voor het meten van de sterkte van een collocatie.
(b) Geef de formule voor het berekenen van Pointwise Mutual Information en vul daarin de juiste waarden in voor het berekenen van de PMI-score van *buitenaards wezen* als:
aantal(*buitenaards*) = 15
aantal(*wezen*) = 10
aantal(*buitenaards wezen*) = 8
corpusgrootte = 1.000.000
Je hoeft de berekening verder zelf niet uit te voeren.
5. Wat wordt bedoeld met de *representativiteit* van een corpus?
6. Wat is het verschil tussen POS-tagging en syntactische annotatie?
7. Leg uit waarom tokenisatie en het splitsen van tekst in zinnen van belang is voor corpustaalkunde.
8. Noem twee voordelen en twee nadelen van het gebruik van Twitterdata in een corpus voor taalkundig onderzoek.
9. Hieronder volgt een lijst met, voor een aantal woord-types in een tekst, de frequentie en de rang (rank) van dit woord-type, wanneer we sorteren op frequentie. In hoeverre bevestigt de lijst de Wet van Zipf?

rank	count
1	1000
2	500
3	333
10	50
25	20
100	5

10. De onderstaande tabel geeft scores voor het voorkomen van de woorden *piano* en *gitaar* met andere woorden in een corpus.

	speel	bespeel	studeer	koop	hoor	steel	stem
piano	100	25	50	100	25	75	300
gitaar	25	200	50	100	50	5	100

Wat is de waarde van Dice(piano,gitaar)? Laat ook zien welke berekening je hebt gebruikt.