Gruppeoppgaver gruppetime IN2110 13.02.2020

1) Skriv opp formelen for Tf-idf og forklar hva hvert element i formelen står for. Hvorfor er det nyttig å bruke Tf-idf?
2) Hva er en centroide? Hvordan finner vi denne? Hva er fordeler og ulemper med å bruke en centroide i motsetning til en medoid?
3) Hva vil det si å lengdenormalisere en vektor? Hvorfor er dette nyttig?
4) I den første obligen lager vi en såkalt Bag of Words (BoW). Hva vil dette si? Nevn noen ulemper ved en slik representasjon.
5) Hvilke to typer læring har vi? Nevn et eksempel på hver.
6) Hvordan fungerer Rochio og hva er noen ulemper med denne?
7) Nedenfor ser dere formelen for å regne ut cosine similarity for to vektorer. Er det i nevneren eller telleren man regner ut lengden til en vektor?

$$sim(A,B) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

8) Under finner du en matrise med termer og kontekster, som inneholder f.eks. TF-IDF eller PPMI verdier for **tre** termer og 5 *kontekst-ord.*
    a. Hva er lengden til ordvektoren for ordet hund?
    b. Hva er cosinus-likheten mellom ordvektoren til hund og ordvektoren til fisk?

	hund	**katt**	**fisk**
løvetann	2	1	3
hund	0	0	1
katt	2	0	1
bil	2	0	2
kopp	2	1	1