# Eksamensopgave 2

## Programmering og sprogteknologi for humanister – ES 2020

For alle opgaverne skal I skrive kommentarer i jeres kode, som kort forklarer, hvad der foregår (dvs. i selve .py/.ipynb-filen skal I skrive små kommentarer). Derudover skal I skrive en rapport, hvor I beskriver jeres program mere detaljeret samt gør rede for jeres valg af løsninger. Skriv på forsiden af jeres aflevering, hvor mange normalsider jeres første rapport var på og hvor mange normalsider den nuværende rapport fylder. I må vedlægge så mange bilag I vil. I bestemmer selv, om I vil aflevere programmerne som .py-filer eller som en (1) .ipynb-fil (men I må ikke aflevere programmerne i et andet format, fx som pdf-fil eller lignende).

Hver opgave er vægtet med et antal point. I skal have min. 55 point for at bestå.

* **Tag skærmprint af en kørsel af hvert enkelt program (dvs. kør programmet og tag et skærmprint af det resultat, som fremkommer).**
* **Vedlæg skærmprint som bilag (også selvom programmet kun virker delvist).**
* **Alle jeres programmer (dvs. selve .py-filerne/ipynb-filen) skal afleveres som bilag.**

Filer du skal bruge til denne opgave: fit\_pos.txt, fit\_neg.txt og stopord.txt. Filerne findes på Absalon. To af filerne indeholder positive og negative anmeldelser af Fitness Word og stopord.txt indeholder ord, som generelt ikke anses som betydningsbærende, og som derfor ofte frasorteres i forbindelse med en kvantitativ undersøgelse (du må meget gerne tilføje ord til/fjerne ord fra stopord.txt, hvis du vil). Anmeldelserne er tilfældigt udvalgt fra Trustpilot, og de repræsenterer ikke nødvendigvis et rimeligt billede af brugeres anmeldelser af denne kæde.

Sammen med anmeldelserne og stopord.txt findes også en fil, som hedder freqout.txt. Denne fil viser et eksempel på, hvordan dit programs output kan se ud. Du skal ikke opfatte den vedlagte freqout.txt som en facitliste – det kan sagtens gøres anderledes, og med andre valg bliver resultatet også anderledes. Men filen kan måske give dig inspiration til, hvordan du synes det bør gøres.

Forestil dig, at du sidder som kommunikationsmedarbejder i Fitness World, og du skal skrive et program, som kan give et overblik over anmeldelserne af FW. Programmet skal bidrage til at udtrække så meget information som muligt fra anmeldelserne og hjælpe dig til at pege på de tendenser, som findes. Det kan være, du så vil kunne give et bidrag til, hvad centeret bør gøre anderledes for at få flere glade kunder, og hvad centeret skal blive ved med at gøre, fordi kunderne allerede er meget glade for det. Det kan være du finder ud af, at kunderne er begejstrede for rengøringen men utilfredse med maskinerne – eller det kan være du finder ud af, at nogle centre får al rosen, mens andre centre får al den hårde kritik. Det kan også være, at du kan se nogle indikationer på, at folk udtrykker sig på en anden måde i en negativ anmeldelse ift. en positiv anmeldelse.

Skriv et program som på forskellig vis foretager en kvantitativ undersøgelse af indholdet i de 2 tekster. Programmet skal være opbygget således, at brugeren kan indtaste 2 filer efter eget valg (det er frivilligt om programmet skal kunne håndtere mere end 2 filer). Dit program skal kunne udføre de undersøgelser, som beskrives i det følgende:

**OPGAVE 1**

1. **Frekvenslister uden at bruge NLTK (25 point**)
2. Dan en frekvensliste over ordene i hver tekst, dvs. der skal være 2 frekvenslister. Hver liste skal være sorteret med det mest frekvente ord øverst.
3. Frekvenslisterne udskrives til freqout.txt. Sørg for, at dit program udskriver overskrifter i freqout.txt, så der findes en overskrift til hver frekvensliste (ellers bliver freqout.txt helt uoverskuelig). Overskriften skal være filnavnet på den fil, frekvenslisten er lavet over.
4. Ord som optræder på stopordslisten skal frasorteres, så disse ord ikke er med på den frekvensliste, som udskrives (du bestemmer selv, om stopord.txt skal bruges som en selvstændig fil - eller om ordene skal indsættes i selve programmet som en liste).
5. Det skal være muligt for brugeren selv at bestemme, hvor mange gange ordet minimum skal optræde i teksten for at blive udskrevet til freqout.txt. Det kan fx være brugeren kun vil se ord, hvor frekvensen er 5 eller højere.
6. **Frekvenslister med NLTK (5 point**)
7. Tag igen udgangspunkt i opgave 1 a., men denne gang skal du indlæse dine filer i NLTK og anvende frekvensliste-faciliteterne (FreqDist) i NLTK.
8. **Bigrammer (10 point)**
9. Dan en frekvensliste for hver tekst over ordpar (bigrammer). Listerne sorteres med de hyppigste bigrammer øverst.
10. Bigramfrekvenslisterne udskrives ligeledes til freqout.txt med passende overskrifter.
11. Giv bruger mulighed for at bestemme, hvor frekvent ordparret skal være for at blive medtaget.
12. Hvis begge ord i et ordpar findes på stopordslisten, skal dette ordpar ikke udskrives til freqout.txt (uanset hvilken frekvens ordparret har).
13. **Fællesmængden af ord (10 point)**
14. Identificer ord, som forekommer i BEGGE filer.
15. Ordene udskrives til freqout.txt.
16. Også her skal det være muligt for brugeren at bestemme, hvor hyppigt ordet skal være for at blive udskrevet til freqout.txt.
17. **Sætningslænger i de to filer (10 point)**

Udskriv til freqout.txt den gennemsnitlige sætningslængde for hver fil. Giv i denne forbindelse en beskrivelse af, hvordan du definerer en sætning og evt. hvilke udfordringer du har set i teksterne ift. definitionen af en sætning.

1. **Særlige ord for de negative/positive anmeldelser (15 point)**
   1. Identificer de ord som findes i fit\_pos.txt og som IKKE findes i fit\_neg.txt (og omvendt: identificer de ord som findes i fit\_neg.txt, men som ikke findes i fit\_pos.txt)
   2. Det skal være muligt for bruger at angive, hvad frekvensen for ordet i fit\_pos.txt (fit\_neg.txt) skal være.
   3. Udskriv resultatet til freqout.txt med passende overskrifter.

Kommunikationen med brugeren i dette program kan se ud på flere forskellige måder. Du bestemmer selv, hvordan kommunikationen skal se ud, men følgende er en mulighed, som du kan overveje at anvende:

Hvilken frekvens skal ordene minimum have i frekvenslisten: 25

Hvilken frekvens skal bigrammerne minimum have:8

Hvilken frekvens skal ordene minimum have i filsammenligningerne:5

Husk at udskrive passende overskrifter i din freqout.txt – så det er nemt at finde rundt i filen (så man fx ved hvilken frekvensliste, der hører til hvilken fil osv.).

**OPGAVE 2 – (10 point)**

Beskriv de tendenser du ser i freqout.txt. Du kan fx forsøge at give et overblik over, hvilke ord, sætningslængder mv., der generelt karakteriserer en positiv anmeldelse ift. en negativ anmeldelse af fitnesscentrene, pege på hvad der især roses eller kritiseres i anmeldelserne samt give din opfattelse af, hvilke af ovenstående undersøgelser, der giver dig mest information mv.

**OPGAVE 3 – (5 point)**

Giv mindst ét forslag (og gerne flere) til, hvordan man kunne forbedre/udvide programmet, så det i endnu højere grad kunne anvendes i en overordnet analyse af anmeldelser.

**OPGAVE 4 – (10 point)**

1. Udskriv (i Jupyter Notebook) et kumulativt frekvensplot over de 30 mest frekvente ord i de to anmeldelsesfiler, før og efter stopord er blevet fjernet.
2. Udskriv de 50 forekomster af ordet ”jeg” i kontekst (konkordans) i a) de negative anmeldelser, b) de positive anmeldelser.
3. Udskriv (i Jupiter Notebook) et dispersion plot for ordene ”jeg”, ”fitness”, ”lokale”,”personale”,”corona” i de to typer anmeldelser, og diskuter hvad de to plots viser.
4. Udskriv ord der optræder i lignende kontekst som ordet ”fitness” i de to anmeldelsesfiler.