Øving 2

Task 1 - Language Model

1. Explain the language model, what are the weaknesses and strengths of this model?

En “language model” er en model som brukes til å forenkle spørringer etter ord i dokumenter. For å sortere etter relevans, brukessannsynligheten og sannsynlighetsdistribusjonen for ord i dokumentet.

“Language model” behandler hvert dokument som et grunnlag for en model. Denne inneholder sannsynligheten for forekomsten av et gitt ord.

+ matematisk presis
+ konseptuelt enkelt
+ intuitiv

- vanskelig å ta høyde for brukerens ønsker
- vanskelig å forbedre relevansen

2. Given the following documents and queries, build the language model according to the document collection.

d1 = *The* apple *is the* pomaceous fruit *of the* apple tree.
d2 = Apple designs *and* creates iPod *and* iTunes. Apple *also* develops Mac operating systems.
d3 = *Is* tomato *a* fruit *or a* vegetable?

d1 = 10 termer
d2 = 13 termer
d3 = 7 termer
Totalt = 30 termer

Use MLE for estimating the unigram model and estimate the query generation probability using the Jelinek-Mercer smoothing

$\hat{P}(t|Md) = (1 - \lambda)\hat{p}mle(t|Md) + \lambda\hat{p}mle(t|C), \lambda = 0.5. (1)$

For each query, rank the documents using the generated scores.

q1 = apple:

P(q1|d1) = ((0.2 + 2/15) * 0.5)) = 0.1666666
P(q1|d2) = ((2/13 + 2/15) * 0.5)) = 0.14358974
P(q1|d3) = ((0 + 2/15) * 0.5)) = 0.06666666666666

Ut ifra queriet og de genererte relevanstallene er rangeringen d1 > d2 > d3

q2 = apple fruit:

P(q2|d1) = ((0.2 + 2/15) * 0.5)) * ((0.1 + 1/15) * 0.5)) = 0.0138888888888888
P(q2|d2) = ((2/13 + 2/15) * 0.5)) * ((0 + 1/15) * 0.5)) = 0.0047863247863247
P(q2|d3) = ((0 + 2/15) * 0.5)) * ((1/7 + 1/15) * 0.5)) = 0.0069841269841269

Ut ifra queriet og de genererte relevanstallene er rangeringen d1 > d3 > d2

q3 = apple coconut

P(q3|d1) = ((0.2 + 2/15) * 0.5)) * ((0 + 0) * 0.5)) = 0
P(q3|d2) = ((2/13 + 2/15) * 0.5)) * ((0 + 0) * 0.5)) = 0
P(q3|d3) = ((0 + 2/15) * 0.5)) * ((0 + 0) * 0.5)) = 0

Ut ifra queriet og de genererte relevanstallene er rangeringen d1=d2=d3

3. Explain what smoothing means and how it affects retrieval scores. Describe your answer using a query from the previous subtask.

Smoothing brukes for å fjerne null-verdier fra søket. Dette gjøres for å få en jevnere model med mindre støy (utligger-verdier). Det skal heller ikke være slik at det er null sannsynlighet for at en ord-sekvens skal forekomme.

Jelinek-Mercer smoothing bruker en lambda verdi (λ) for å avgjøre en querie terms innvirkning på søket.
Et eksempel på dette er q3.

Task 2 - Evaluation of IR Systems

1. Explain the terms Precision and Recall, including their formulas. Describe how differently these metrics can evaluate the retrieval quality of an IR system.

Precision er andelen av de uthentede dokumentene som er relevante.
Precision = (relevant items retrieved) / (retrieved items) = P(relevant|retrieved)

Recall er andelen av de relevante dokumentene som er hentet ut.
Recall = (relevant items retrieved) / (relevant items) = P(retrieved|relevant)

For å evaluere kvaliteten på disse metrikkene bruker vi følgende tabell

	Relevant	Nonrelevant
Retrieved	True positives (TP)	False positives (FP)
Not retrieved	False positives (FN)	True negatives (TN)

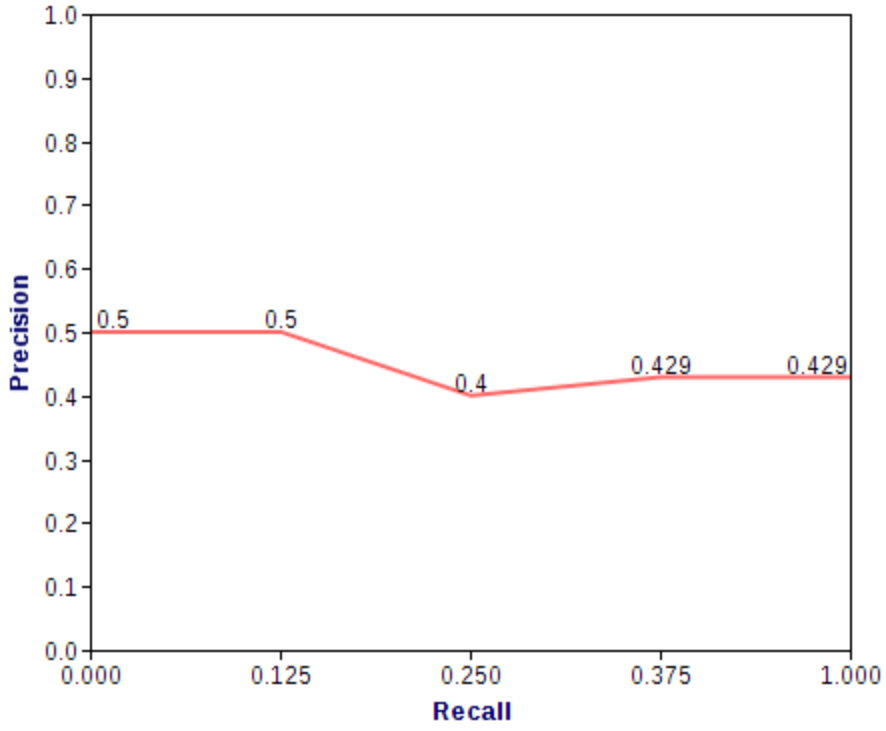
Vi ser da at formelene for kvalitet i Precision og Recall er
Precision = TP/(TP + FP)
Recall = TP/(TP + FN)

2. Given the following set of relevant documents

Given the following set of relevant documents
rel = {201, 6, 72, 10, 84, 15, 92, 33, 37, 46}, and the set of retrieved documents
ret = {60, 92, 103, 2, 201, 66, 33, 45}, provide a table with the calculated Precision and Recall at each level.

d	**Relevant**	**Precision**	**Recall**
60			
92	REL	0.5	0.1
103			
2			
201	REL	0.4	0.2

66			
33	REL	0.428	0.3
45			



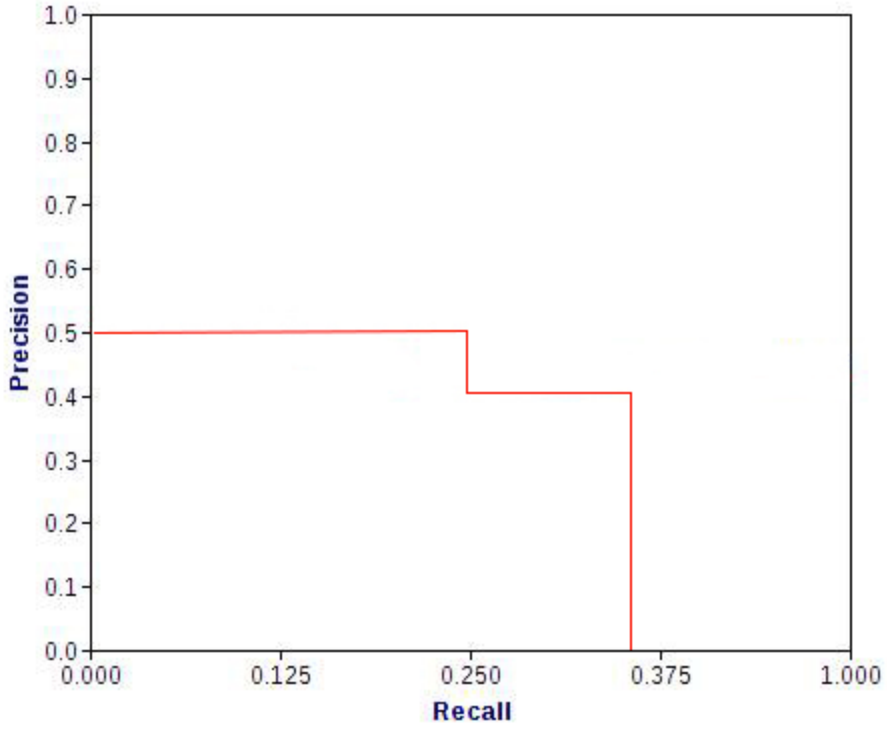
Task 3 - Interpolated Precision

1. What is interpolated precision?

I arrangert gjenfinning har man recall og precision. Der recall er andelen av de relevante dokumentene som er hentet, er precision hvor nøyaktig systemet har vært til å hente ut dokumenter.

Interpolated precision er når du velger en recall-verdi r og for alle recall-verdier r' >= r, er det den beste precision du kan få.

2. Given the example in Task 2.2, find the interpolated precision and make a graph.



Task 4 - Relevance Feedback

1. What is the purpose of relevance feedback? Explain the terms Query Expansion and Term Re-weighting. What separates the two?

Hensikten med relevans feedback er er å bruke et resultat fra et gitt query, og bruke det til å finne ut om det er relevant å bruke til å utføre et nytt query
Query Expansion er å evaluere brukerens input og utvidet queriet til å treffe flere dokumenter. *Term Re-weighting* handler om å gjøre vekten på termer i urelevante dokumenter mindre, og øke vekten på termer i relevante dokumenter.
Forskjellen på *Query Expansion* og *Term Re-weighting* er at reweighting ikke bruker spørringsutvidelse

2. Explain the difference between automatic local analysis and automatic global analysis.

I automatic local analysis blir kun dokumenter som blir returnert av queriet evaluert, dette blir gjort uten hjelp fra bruker. Global feedback bruker informasjon fra dokumenter i hele kolleksjonen.