An Example of Language Models for IR

Dell Zhang
16/11/2008

Query

$q$: “tail head tail head tail tail”

Document Collection

$d_1$: “head head head tail head head.”
$d_2$: “tail tail head tail head head.”
$d_3$: “tail head tail tail tail head.”

How shall we rank the documents w.r.t. the query using document unigram models (without smoothing)?
Constructing the Unigram Language Model

\[ d_1 \Rightarrow M_1: \quad P(\text{head}|M_1) = 5/6 \quad P(\text{tail}|M_1) = 1/6 \]

\[ d_2 \Rightarrow M_2: \quad P(\text{head}|M_2) = 1/2 \quad P(\text{tail}|M_2) = 1/2 \]

\[ d_3 \Rightarrow M_3: \quad P(\text{head}|M_3) = 1/3 \quad P(\text{tail}|M_3) = 2/3 \]

Applying the Unigram Language Model

\[
P(q|M_1) = P(\text{"tail head tail head tail tail"}|M_1) = P(\text{tail}|M_1) P(\text{head}|M_1) P(\text{tail}|M_1) P(\text{head}|M_1) P(\text{tail}|M_1)
\]

\[
P(\text{tail}|M_1) = (1/6) * (5/6) * (1/6) * (5/6) * (1/6) \approx 0.0005
\]

\[
P(q|M_2) = P(\text{"tail head tail head tail tail"}|M_2) = P(\text{tail}|M_2) P(\text{head}|M_2) P(\text{tail}|M_2) P(\text{head}|M_2) P(\text{tail}|M_2)
\]

\[
P(\text{tail}|M_2) = (1/2) * (1/2) * (1/2) * (1/2) * (1/2) \approx 0.0156
\]

\[
P(q|M_3) = P(\text{"tail head tail head tail tail"}|M_3) = P(\text{tail}|M_3) P(\text{head}|M_3) P(\text{tail}|M_3) P(\text{head}|M_3) P(\text{tail}|M_3)
\]

\[
P(\text{tail}|M_3) = (2/3) * (1/3) * (2/3) * (1/3) * (2/3) \approx 0.0219
\]

Probabilistic Ranking Principle

The returned list of documents should be in the order of \(d_3, d_2, d_1\). because \(P(q|M_3) > P(q|M_2) > P(q|M_1)\).