Language Modelling for IR

- A common suggestion to users for coming up with good queries is to think of words that would likely appear in a relevant document, and to use those words as the query.
- The language modeling approach to IR directly models that idea: a document is a good match to a query if the document model is likely to generate the query, which will in turn happen if the document contains the query
- This approach thus provides a different realization of some of the basic ideas for document ranking.
- Language modeling — that is, predicting the probability of a word in a sentence — is a fundamental task in natural language processing. It is used in many NLP applications such as autocomplete, spelling correction, or text generation.

- Instead of overtly modeling the probability $p(R=1|q,d)$ of relevance of a document d to a query q, as in the traditional probabilistic approach to IR.
- The basic language modeling approach instead builds a probabilistic language model M_d from each document d, and ranks documents based on the probability of the model generating the query: $P(q|M_d)$

- Simplest form of language model is known as unigram language model, is a probability distribution over the words in the language.
- For example, if the document in a collection contained five different words, a possible language model for that collection might be (0.2, 0.1, 0.35, 0.25, 0.1), where each number is the probability of a word occurring. If we treat each document as a sequence of words, then the probabilities in the language model predicts what the next word in the sequence will be.
- let the five words be “cat”, “rain”, “dog”, “jump”, “the”

- With this model, for example, it is just likely to get the sequence “cat rain” (0.2 * 0.1) as “cat the” (0.2 * 0.1)
- In n gram language model, prediction of words based on longer sequences are used. An n-gram model predicts a word on the previous (n-1) words.