Exercise 9

Information Retrieval

Dresden Database Systems Group

12. Language Models for IR

TECHNISCHE UNIVERSITÄT DRESDEN

Warm up

ok

Exercise 12.1

- Are the following statements true or false? Give reasons for your answer.

a) A statistical language model (LM) assigns probabilities to strings of symbols from some alphabet.

b) For every unigram LM, `cats hunt mice` and `mice hunt cats` have the same probability.

c) Language models are completely unrelated to Markov chains.

d) We use LMs in IR like this: (step 1) we derive a LM from each document, (step 2) we rank all documents by the probability that their LM generates the query.

e) From a vocabulary containing $|V|$ terms we can construct approximately $|V|^n$ n-grams.

f) A document containing $|d|$ tokens contains approximately $|d|^n$ n-grams.

g) In practice, language models in IR consider n-grams with $n \geq 3$, i.e., at least tri-grams.

h) The shorter the query, the more important is smoothing.

IR Using a Unigram Language Model

Exercise 12.2

- The table below provides information on a corpus of three documents
- We focus on the terms `tasty`, `coffee`, and `sugar` (but there are more terms in the vocabulary)

Document d	$\lvert d \rvert$	Term frequencies $tf_{d,t}$		
		`tasty`	`coffee`	`sugar`
D1	200	20	100	20
D2	100	0	10	0
D3	200	0	40	20

根据贝叶斯公式推导的

- A user submits the query `tasty coffee tasty sugar`

- Calculate the documents' scores using a unigram LM and determine the ranking
  a) without any smoothing
  b) using Jelinek-Mercer-Smoothing (with $\lambda = \frac{1}{2}$) 画表格，需要再熟练！形成做题讲课规范 next