信息检索大作业

Zhenghao Liu　THUNLP
liuzhenghao0819@163.com

Outline

- Preliminary
  - Traditional IR Models
- Goal of This Project
- Experimental Setting
- Present and Report

Traditional IR Models

- Traditional IR model
  - The language modeling approach to IR is quite extensible:

$$\mathrm{p}(d|q) \approx p(q|d)p(d)$$

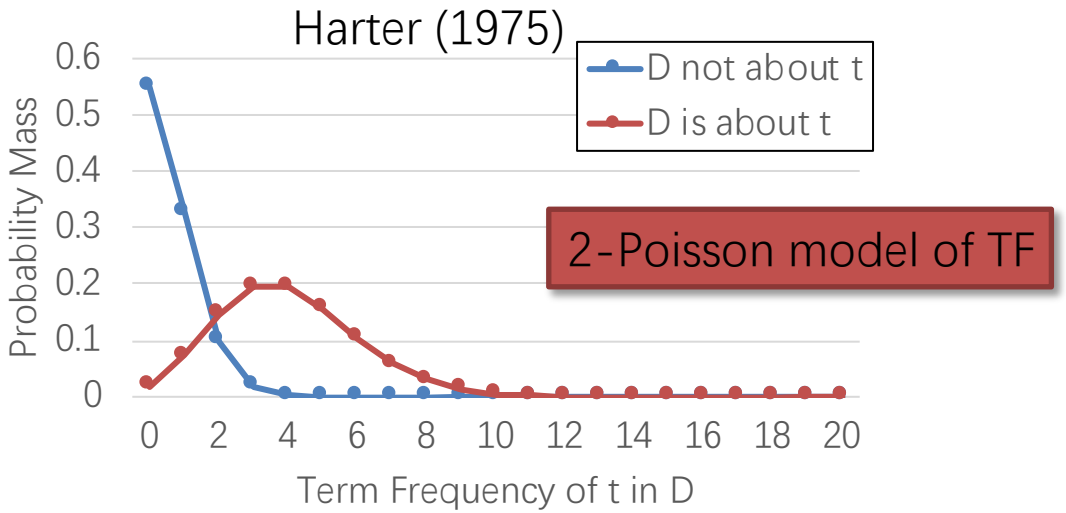
    - $\mathrm{p}(d)$ can be assumed uniform across docs
    - $\mathrm{p}(q|d) = \prod_{w \in q} p(w|d)$ depends on how to model the relationship of query word and doc

Traditional IR Models

- TF-IDF
  - Term Frequency (TF)
    - The number of times that term t occurs in document d:

$$tf(t, D) = \frac{n_t}{n_d}$$

      - Where n_t is the number of times the word t appears in d, and n_d is the word number of the document (d)



Traditional IR Models

- TF-IDF
  - Term Frequency (TF)
  - Inverse Document Frequency (IDF)
    - IDF is a measure to evaluate if term t is common or rare across the document collection

$$\mathrm{IDF}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

    - TF-IDF

$$\mathrm{TF-IDF}(t, D) = \mathrm{TF}(t, D) \cdot \mathrm{IDF}(\mathrm{t}, \mathrm{D})$$

Traditional IR Models

- BM25
  - BM25 is a bag-of-word retrieval model
    - Given a query Q, which contains n words q_1, ...q_n, the BM25 score of a document D is:

$$score(D,Q) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{f(q_i,D) \cdot (k+1)}{f(q_i,D) + k \cdot \left(1 - b + b \cdot \frac{|D|}{avgdl}\right)}$$

      - Where $f(q_i, D)$ is the term frequency of q_i in the document D, $|D|$ is the length of D, and $avgdl$ is the average document length in the document collection
    - BM25 aims to normalize term frequency according to document length

Traditional IR Models

- BM25
  - BM25 is a bag-of-words retrieval

$$score(D,Q) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{f(q_i,D) \cdot (k+1)}{f(q_i,D) + k \cdot \left(1 - b + b \cdot \frac{|D|}{avgdl}\right)}$$

    - k and b are free parameters:
      - If k is large enough, $\frac{f(q_i,D)\cdot(k+1)}{f(q_i,D)+k\cdot\left(1-b+b\cdot\frac{|D|}{avgdl}\right)} \approx f(q_i,D)$
      - If $k = 0$, $\frac{f(q_i,D)\cdot(k+1)}{f(q_i,D)+k\cdot\left(1-b+b\cdot\frac{|D|}{avgdl}\right)} = 1$
      - $0 \leq b \leq 1$:
        - » If $b = 0$, we do not consider document length
        - » If $b = 1$, we normalize term frequency totally according to document length

Traditional IR Models

- Neural Translation Language Model (NTLM)
  - Translation Language Model: extend query likelihood: $p(d|q) \sim p(q|d)p(d)$

$$p(q|d) = \prod_{t_q \in q} p(t_q|d)$$

$$p(t_q|d) = \sum_{t_d \in d} p(t_q|t_d)p(t_d|d)$$

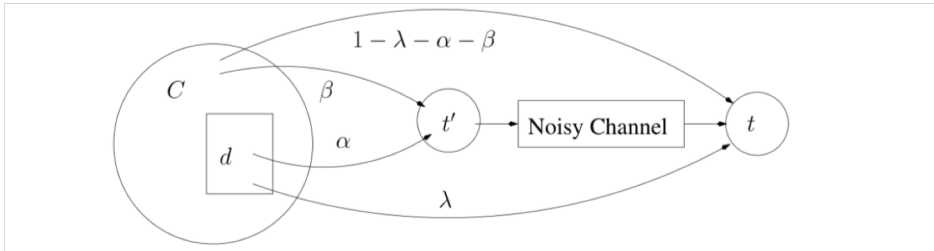
  - Use the similarity between term embeddings as a measure for term-term translation probability $p(t_q|t_d)$

$$p(t_q|t_d) = \frac{\cos(\vec{v}_{t_q}, \vec{v}_{t_d})}{\sum_{t \in V} \cos(\vec{v}_t, \vec{v}_{t_d})}$$

Berger and Lafferty, "Information retrieval as statistical translation." SIGIR 1999

Traditional IR Models

- Generalize Language Model (GLM):
  - Term t in a query is generated by sampling independently from either the document or the document collection



$$p(t|d) = \lambda p(t|d) + \alpha \sum_{t' \in d} p(t\,|t')p(t'|d) + \beta \sum_{t' \in N_t} p(t|t')p(t'|C) + (1-\lambda-\alpha-\beta)\, p(t\,|C)$$

  - The noisy channel may transform (mutate) a term t' into a term t. Term t'' is sampled from its nearest neighbors

$$p(t|t') = \frac{sim(\vec{v}_{t'}, \vec{v}_t)}{\sum sim(\vec{v}_{t'}, \vec{v}_{t''})}$$

Ganguly, Roy, Mitra, and Jones, Word embedding based generalized language model for information retrieval. SIGIR 2015

Outline

- Preliminary
- Goal of This Project
- Experimental Setting
- Demo and Report

Goal of This Project

- 实现一个完整的信息检索系统并达到以下目的
  - 对大规模中文文本分词，词性标注
  - 给定关键词，查询出相对应的常见搭配，并考虑如下场景：
    - 返回关键词常见搭配的结果（列表）
    - 给定返回关键词词性，返回常见搭配的结果
    - 给定检索词与关键词距离限制，返回常见搭配的结果（窗口）
    - 给定多个关键词返回相应的结果
    - 分词或不分词对检索效果的影响
  - 最终形成报告以及展示系统

Outline

- Preliminary
- Goal of This Project
- Experimental Setting
- Present and Report

Experimental Setting

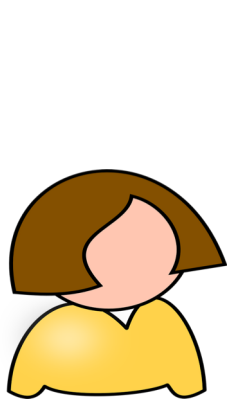
- 分词工具：THULAC
  - http://thulac.thunlp.org
- 索引搭建： Lucene (Pylucene, Elasticsearch)
  - Pylucene 参考链接
    - https://github.com/hasibi/EntityLinkingRetrieval-ELR
- 编程语言：Python
- Demo实现：Flask
- 语料库：Sougou-T、中文Wiki以及人民日报语料
  - Sougou-T以及人民日报下载链接（注意：不可公开）
    - https://cloud.tsinghua.edu.cn/d/77df14a5af484eb685b9/
  - 分组要求：
    - 1人一组

Outline

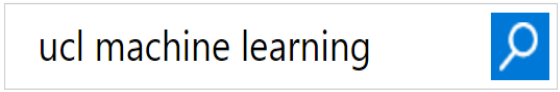
- Preliminary
- Goal of This Project
- Experimental Setting
- Present and Report

Present and Report

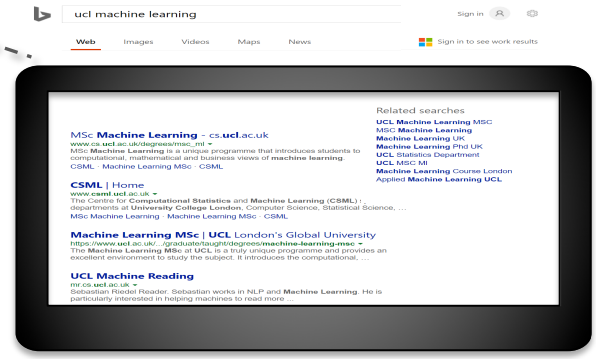
- Demo



Relevance
(key words)



words



Results (keywords list)

retrieval
system indexes
a document
corpus

Present and Report

- 报告提交：PDF
- 代码提交：Readme，索引构建以及Demo代码以及必要的注释
- 打分细则：
  - 报告30%、功能40%、语料库规模30%