SIGNIFICANT WORDS REPRESENTATIONS OF ENTITIES
LUHN-BASED MODEL

- How to estimate a model for representing a set of textual entities capturing all, and only, the essential shared commonalities of these entities?

Revisiting Luhn!

Significant Words Language Models

HOW TO ESTIMATE SWLM?

$p(t|d) = \lambda_{d,sw}p(t|\theta_{sw}) + \lambda_{d,g}p(t|\theta_g) + \lambda_{d,s}p(t|\theta_s)$

$p(t|\theta_g) = p(t|\theta_C) = \frac{c(t,C)}{\sum_{t'\in V} c(t',C)}$

Probability of term $t$ to be important in one of the document models but not others, marginalizing over all the documents.

$p(t|\theta_s) = \sum_{d_i \in D} \left( p(t|\theta_{d_i}) \prod_{d_j \in D, j \neq i} (1 - p(t|\theta_{d_j})) \right)$

Latent Variables:

$\Upsilon = \{\lambda_{d,sw}, \lambda_{d,g}, \lambda_{d,s}\}_{d \in D} \cup \{\theta_{sw}\}$

Fit the log-likelihood model of generating all terms in the documents in the set:

$log p(D|\Upsilon) = \sum_{d \in D} \sum_{t \in V} c(t,d) \log \left( \sum_{x \in \{sw,g,s\}} \lambda_{d,x} p(t|x) \right)$

Solve this problem:

$\Upsilon^* = \arg\max_{\Upsilon} p(D|\Upsilon)$

E-Step:

$p(X_{d,t} = x) = \frac{p(\theta_x|d)p(t|\theta_x)}{\sum_{x' \in \{sw,g,s\}} p(\theta_{x'}|d)p(t|\theta_{x'})}$

M-Step:

$p(t|\theta_{sw}) = \frac{\sum_{d \in D} c(t,d)p(X_{d,t} = r)}{\sum_{t' \in V} \sum_{d \in D} c(t',d)p(X_{d,t'} = r)}$

$\lambda_{d,x} = p(\theta_x|d) = \frac{\sum_{t \in V} c(t,d)p(X_{d,t} = x)}{\sum_{x' \in \{sw,g,s\}} \sum_{t \in V} c(t,d)p(X_{d,t} = x')}$
Group Profiling: To understand and model the characteristics of the group of entities.

Content Customisation: Contextual Suggestion

Performance of employing user preferences-based and group-based customization on contextual suggestion task.
SWLM APPLICATIONS: REGULARIZED SWLM

- **Relevance Feedback**: Capturing the “mutual notion of relevance”:
  - **RSWLM** incorporates the extra knowledge from the query model by defining a prior parameter

\[
p(\theta_{sw}) \propto \prod_{t \in V} p(t|\theta_{sw})^{\beta} p(t|\theta_q)
\]

- Topic 374 of the TREC Robust04 test collection: `Nobel prize winners`

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>prize 5.56e-02</td>
<td>new 3.70e-03</td>
<td>prize 6.07e-02</td>
<td>insulin 2.25e-02</td>
<td>prize 6.02e-02</td>
</tr>
<tr>
<td>nobel 3.36e-02</td>
<td>cent 2.98e-03</td>
<td>nobel 4.37e-02</td>
<td>palestinian 2.15e-02</td>
<td>nobel 4.53e-02</td>
</tr>
<tr>
<td>physics 2.35e-02</td>
<td>two 2.97e-03</td>
<td>awards 3.43e-02</td>
<td>dehmelt 1.81e-02</td>
<td>science 2.68e-02</td>
</tr>
<tr>
<td>science 2.18e-02</td>
<td>dollars 2.76e-03</td>
<td>chemistry 3.23e-02</td>
<td>oscillations 1.79e-02</td>
<td>award 2.43e-02</td>
</tr>
<tr>
<td>time 1.68e-02</td>
<td>people 2.71e-03</td>
<td>physics 2.82e-02</td>
<td>waxman 1.69e-02</td>
<td>physics 1.94e-02</td>
</tr>
<tr>
<td>palestinian 1.34e-02</td>
<td>...</td>
<td>palestinian 2.18e-02</td>
<td>marcus 1.69e-02</td>
<td>winner 1.90e-02</td>
</tr>
<tr>
<td>year 1.34e-02</td>
<td>...</td>
<td>cesium 2.09e-02</td>
<td>attack 1.61e-02</td>
<td>won 1.80e-02</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>arafat 1.94e-02</td>
<td>...</td>
<td>peace 1.80e-02</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>university 1.92e-02</td>
<td>...</td>
<td>discovery 1.71e-02</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>arafat 1.29e-02</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
SWLM APPLICATIONS: REGULARIZED SWLM

- Dealing with poison pills: (Harmful relevant documents)

Effectiveness of different feedback systems facing with bad relevant document in topic 374 of TREC Robust04
**Modelling Hierarchical Entities**

\[
p(t|d) = \lambda_{d,g} p(t|\theta_g) + \lambda_{d,sw} p(t|\theta_{sw}) + \lambda_d s p(t|\theta_s)
\]
**SWLM APPLICATIONS: HIERARCHICAL SWLM**

- **Hierarchical Classification:**
  - **Two-dimensionally separable**
    - Transferable models for classification of entities in *evolving* hierarchies
SWLM AS AN ANALYTICAL TOOL

Score Decomposition:

- Dynamically determining the contribution of each document

![Diagram](Image.png)

(a) Robust04 dataset  
(b) WT10G dataset  
(c) GOV2 dataset
DON'T BUILD YOUR MODEL BASED ON PROPERTYLESS COMMON OBSERVATIONS OR UNRELIABLE RARE OBSERVATIONS, TAKE THE SIGNIFICANT ONES.

Luhn and Mostafa!