

# ML Week

## Natural Language Processing

Jeff Abrahamson

23–25 novembre 2016

$$\begin{array}{ll}\text{Maximize} & c^T x \\ \text{subject to} & Ax \leq b\end{array}$$

# Summarising Text

- Abstractive (hard)
- Extractive (select sentences)

# Summarising Text

Challenge problem (cf. greedy solutions):

The cat is in the kitchen.

The cat drinks the milk.

The cat drinks the milk in the kitchen.

# Summarising Text

- Sentence selection
- Use n-grams
- Stemming
- Stop words
- Prune short sentences

*Dan Gillick, Benoit Favre, A Scalable Global Model for Summarization, 2009*

# Summarising Text

## Outline:

- ILP (*optimisation linéaire en nombres entiers*)
- Maximum coverage model

*Dan Gillick, Benoit Favre, A Scalable Global Model for Summarization, 2009*

# Summarising Text

ILP in canonical form:

$$\begin{array}{ll}\text{Maximize} & c^T x \\ \text{subject to} & Ax \leq b \\ & x \geq 0 \\ & x \in \mathbb{Z}^n\end{array}$$

*Dan Gillick, Benoit Favre, A Scalable Global Model for Summarization, 2009*

# Summarising Text

ILP in standard form:

$$\begin{array}{ll}\text{Maximize} & c^T x \\ \text{subject to} & Ax + s = b \\ & s \geq 0 \\ & x \in \mathbb{Z}^n\end{array}$$

*Dan Gillick, Benoit Favre, A Scalable Global Model for Summarization, 2009*



# Summarising Text

ILP in standard form:

$$\begin{array}{ll}\text{Maximize} & c^T x \\ \text{subject to} & Ax + s = b \\ & s \geq 0 \\ & x \in \mathbb{Z}^n\end{array}$$

This is NP hard.

*Dan Gillick, Benoit Favre, A Scalable Global Model for Summarization, 2009*

# Summarising Text

ILP in standard form:

$$\begin{array}{ll}\text{Maximize} & c^T x \\ \text{subject to} & Ax + s = b \\ & s \geq 0 \\ & x \in \mathbb{Z}^n\end{array}$$

Discussion: linear vs integer programming.

*Dan Gillick, Benoit Favre, A Scalable Global Model for Summarization, 2009*

# Summarising Text

Let

$c_i$  : presence of concept  $i$  in summary

$w_i$  : weight associated with  $c_i$

$l_i$  : length of sentence  $i$

$s_j$  : presence of sentence  $j$  in summary

$L$  : summary length limit

$Occ_{ij}$  : occurrence of  $c_i$  in  $s_j$

*Dan Gillick, Benoit Favre, A Scalable Global Model for Summarization, 2009*

# Summarising Text

## Summarisation

$$\begin{aligned} & \text{Maximize} && \sum_i w_i c_i \\ & \text{subject to} && \sum_j l_j s_j \leq L \\ & && s_j \text{Occ}_{ij} \leq c_i, && \forall i, j \\ & && \sum_j s_j \text{Occ}_{ij} \geq c_i && \forall i \\ & && c_j \in \{0, 1\}, && \forall j \\ & && s_j \in \{0, 1\}, && \forall j \end{aligned}$$

# Summarising Text

## Notes:

- Selecting a sentence selects all concepts it contains
- Selecting a concept requires it be in at least one sentence
- $s_j Occ_{ij} \leq c_i, \forall i, j \Rightarrow$  no concept-less sentences

*Dan Gillick, Benoit Favre, A Scalable Global Model for Summarization, 2009*

# Sentiment Analysis

Many variations:

- Entire documents using computational linguistics
- Manually crafted lexicons

# Sentiment Analysis

## Techniques

- Template instantiation (requires domain knowledge)
- Passage extraction

# Sentiment Analysis

- Extract “opinion sentences” based on the presence of a predetermined list of product features and adjectives.
- Evaluate the sentences based on counts of positive vs negative polarity words (as determined by the Wordnet algorithm)

*Hu and Lieu, Mining and Summarizing Customer Reviews, 2004*



# Sentiment Analysis

- Extract “opinion sentences” based on the presence of a predetermined list of product features and adjectives.
  - “The food is excellent.”
  - “The food is an excellent example of how not to cook.”
- Evaluate the sentences based on counts of positive vs negative polarity words (as determined by the Wordnet algorithm)

*Hu and Lieu, Mining and Summarizing Customer Reviews, 2004*

# Sentiment Analysis

- Extract “opinion sentences” based on the presence of a predetermined list of product features and adjectives.
- Evaluate the sentences based on counts of positive vs negative polarity words (as determined by the Wordnet algorithm)

The good: fast, no training data, decent prediction.

The bad: fails on multiple word sense, non-adjectives; sensitive to context.

*Hu and Lieu, Mining and Summarizing Customer Reviews, 2004*

# Sentiment Analysis

Words aren't enough.

- “unpredictable plot” vs “unpredictable performance”

*Turney, Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews, 2002*

# Questions?

[ml-week.com/1](http://ml-week.com/1)