

# On-line summarisation of time-series documents

Ersi Ni

Supervisor: Dr. Dong, Prof. Dr. Smyth  
School of Computer Science  
University College Dublin

Presentation of FYP, 2017

# Agenda

1. Introduction of the problem
2. Solution proposal and challenges
3. Solution stage 1: exploratory analysis on time-series
4. Solution stage 2: natural language processing
5. Result with UI Mock-Up

# A story about hotels (in i.e. Dublin)

1. User side: Hotel booking
2. Hotel side: Feedback as reasoning for business decisions

# Story ctd.

- \* Core problem: information overload (reviews)

For Users:

1. Rating is obvious, but 5 hotels with similar price and similar rating scores, which to choose?
2. What other users are saying about this hotel?

For Hotels:

1. Trending topic (context): Why rating has gone down for the past 3 weeks
2. Background from the city: Revealing what (periodical) events is happening in the past
3. Users have chosen us in this August over our competitors because...?

# Solution Proposal

1. Using statistics from activities in time periods and rating scores to pin-point representative context.
2. Summarise the context as information retrieval.

# Challenges

1. Data source
  - 1.1 availability
  - 1.2 feature set
  - 1.3 consistency
  - 1.4 technical challenges
2. Dimension of the data
3. Need exploration of the data before asking the real question (Exploratory Analysis)
4. Text Summarisation (pretty much open topic)

Excerpt of the statistics

feature	count
word tokens	17,471,927
sentences	1,000,631
reviews	200,738
hotels	700 (Dublin, Galway and Cork)

# Data acquisition

1. Scrapping from web
  - 1.1 define feature
  - 1.2 parsing
  - 1.3 irregularity handling
2. Normalisation and Output:
  - 2.1 Hotel set [meta data]
  - 2.2 Review set [meta data AND review text]

# Exploratory Analysis on time-series

Hotel is static data consisting of meta data, but reviews are time-series.

Top	First	Last
2015-05-26	2001-11-28	2016-09-19

Figure: Timestamp statistics

Example exploratory: what can rating scores and count tells us about one hotel for the past 3 years?

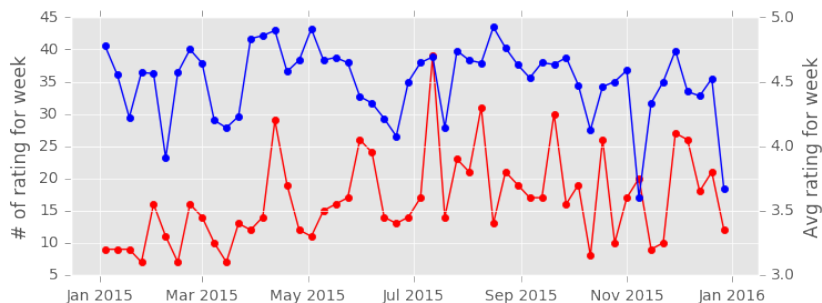
This requires further normalisation:

1. Subset to 3 years (then by hotel)
2. And resample to 365 days each year
3. Aggregation statistics on periods like week, month, 3 months etc.



## Exploratory Analysis ctd.

Gibson Hotel from 2015 to 2016 (subset to 1 year to fit in this slide)



**Figure:** Gibson Hotel aggregated mean rating score and count by week, for 1 year. Red: Count, Blue: average rating score.

## Stage 2: Text Summarisation

Assume that we want to investigate the week starting on Sunday the 2015 November 8th for Gibson Hotel.

( the sudden drop in average rating score to mere 3.6 )

We look at the 13 reviews for Gibson in that week and compare to 869 reviews for all hotels.

Strategy for summarisation:

Construct a  $n$  sentences long summary by **extracting  $n$  *key*** sentences from the text pool.

# Graph based summarisation

**Key** sentences are defined by prestige (importance, uniqueness)

1. Word centrality is defined by a score called  $tf \times idf$ .
2. Sentences are presented as graph nodes. Degrees between the nodes are voted on a similarity score from a modified formula based on the said centrality.
3. More connections between nodes means more prestige.

Inspired by Google's Pagerank algorithm, this lexicon modified version called **LexRank** is introduced by [Erkan & Radev 2004].

Visualise the effect of  $tf \times idf$  on words.



Figure: Word Cloud constructed based on  $tf \times idf$  of the reviews for Dublin (left) and Gibson Hotel (right) in the week from 2015-11-08 to 2015-11-15

## Result summarisation with UI Mock-Up

Overview

Rooms & Rates

Reviews (4,389)

Photos (1,720)

Location

Amenities

executive room

spacious rooms

would highly recommend this hotel

following morning

enjoyable stay

arena

docklands

liffey

iron

Traveller rating

☐ Excellent

2,599

☐ Very good

1,037

☐ Average

250

☐ Poor

86

☐ Terrible

48

Traveller type

☐ Families (501)

☐ Couples (1,811)

☐ Solo (132)

☐ Business (563)

☐ Friends (657)

Time of year

☐ Mar-May (969)

☐ Jun-Aug (1,125)

☐ Sep-Nov (1,064)

☐ Dec-Feb (862)

Language

☐ All languages

☒ English (4,020)

☐ Italian (99)

☐ German (95)


More

Showing 4,020: English reviews


Clear all

Base on our analysis of what people said in the past, this is a summarisation related to the similar period of your intended stay:

1. The reason for the stay was due to Micheal McIntyre show on in the 3arena.
2. We went out for meal the first night, and ate in the hotel on the second night.
3. The Gibson hotel is located literally a stones throw from the 3 arena which makes it a perfect place to stay if you are going to see someone play here.
4. The hotel does not have it's own car park so you have to park in the Point Village car park which connects to the hotel.
5. This is a convenient hotel for the 3 arena venue but certainly not somewhere for a quiet night away.

 **wynonahclaire**  
Austin, Texas

*“Good location, nice hotel, don't order pizza”*

 Reviewed 2 days ago

We love how environmentally friendly The Gibson is! Very cool location and rooms are nice and large. The room service was friendly and quick

# Achieved outcome recap

## 1. Software Pipeline

- 1.1 Data acquisition
- 1.2 Software module for time-series manipulation
- 1.3 Software implementation of NLP modules for Graph-based solution
- 1.4 Software implementation for **LexRank** specifically for multi-document summarisation (different to existing modules available online)

## 2. Dataset for future research

- 2.1 Hotel and Review meta data as **csv** files
- 2.2 Raw review text for NLP as compressed archive
- 2.3 Genre (Hotel reviews) wide dataset for NLP [tokens, sentences, term-frequency, inverse-document-frequency etc.]

# Outcome vs original project spec

## Difference

1. Hotel review as target domain instead of international news
2. Produced dataset for future research

## Matched

1. Novelty through combining event detection with existing NLP summarisation
2. "Live" data summarisation is achievable using cached corpus
3. Chronology is preserved

## Pending

1. Evaluation against other time-series summarisation