THE UNIVERSITY OF MELBOURNE

Workshop 5

COMP20008

Elements of Data Processing

Zijie Xu

Agenda

- N-grams and text similarity
- Text pre-processing and Bag of Words
- Data visualisation

N-grams and text similarity

- N-gram: sequences of n contiguous words within text
- Letter N-gram: substrings of n contiguous letters within string
  - Bi-grams of ‘crat’: $G_2(\texttt{crat}) = [\texttt{\#c, cr, ra, at, t\#}]$
  - ’#’ stands for padding
- We can use n-gram/letter n-gram as basic units to compare how similar two documents/strings are

N-grams and text similarity

- There are many metrics for finding the similarity between two strings/documents
  - N-gram: Simple n-gram, Jaccard, Dice
  - Edit distance based: Levenshtein, Hamming
  - Geometric: Cosine
  - Distance based: Manhattan, Euclidean, Minkowski

Text pre-processing

1. Sentence splitting
   - Goal: split word strings into sentences
   - Problem: Numbers ($3.50) and abbreviations (a.k.a.)
2. Word splitting
   - Goal: split sentences into word tokens
   - Problem: punctuations (it's), cases (US & us) and other non-canonical text forms ($1,234.56 & €1.234,56 13:30 & 1.30 pm)
   - Tool: regular expression

Text pre-processing

3. Word regularisation
   - Goal: find one representation for many morphologies of a word
   - Problem: number, tense, aspect; root + more parts
   - Tools: stemming (Porter stemmer), Lemmatising (WordNet)
4. Bag of words
   - Representing a document as a count of every word it contains
   - N-gram: N successive word(s) as a unit
   - One way to represent text as a numeric vector

Data visualisation

- Some common charts
  - **Histograms** Represent the distribution of univariate data
  - **Boxplot** Summarise the quartiles of univariate data
  - **Bar Charts** Used to compare values across categories
  - **Scatter Plots** Display relationships between two variables
  - **Heatmaps** Represent data using colours on a grid
- Both Matplotlib and Seaborn are great packages to plot visualisation with python
  - `import matplotlib.pyplot as plt`
  - `import seaborn as sns`

THE UNIVERSITY OF MELBOURNE

Thank you

More Resources: Canvas