Data Mining

UNIVERSITY OF LEEDS

In this lecture you will be introduced to

- lecturer: Eric Atwell
- Overview of Data Mining module
- The assessment for this module
- background and practical applications of data mining and text analytics

Texts you should browse:

- E Atwell. 1999. The language machine. British Council.
- https://eprints.whiterose.ac.uk/81779/1/TheLanguageMachine.pdf
- Assessment specification

Lecturer: Eric Atwell

UNIVERSITY OF LEEDS

Professor of Artificial Intelligence for Language

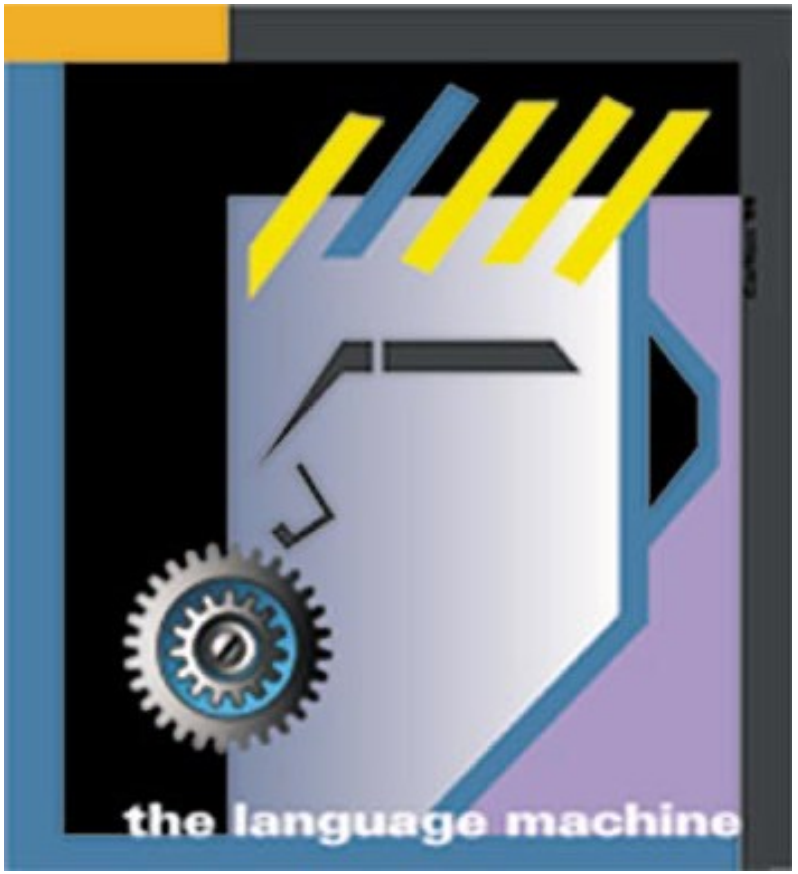
www.comp.leeds.ac.uk/eric

Research includes:

Religious Text Analytics

Arabic Corpus Linguistics

Chatbots for Education



Data Mining and Text Analytics module overview (i)

UNIVERSITY OF LEEDS

01 intro.pptx

02a text and words RegEx corpora SLP.pptx

02b SketchEngine.pptx

03a ngram language models SLP.pptx

03b data text social media.pptx

04a text classifiers sentiment evaluation SLP.pptx

04b scaling to big data.pptx

05a word meanings embeddings SLP.pptx

05b writing project proposal.pptx

Data Mining and Text Analytics module overview (ii)

UNIVERSITY OF LEEDS

06a tagging POS and NER.pptx

06b Machine Translation.pptx

07a Information Extraction.pptx

07b CHEAT NLTK Python.pptx

08a Unsupervised Machine Learning.pptx

08b Information Retrieval.pptx

09a chatbots dialogue SLP.pptx

09b BERT Large Language Model.pptx

10a edubots for university education.pdf

PLUS: extension activities…

Textbook

UNIVERSITY OF LEEDS

Jurafsky, D., & Martin, J. (forthcoming).

Speech and Language Processing, 3rd edition. Pearson See: https://web.stanford.edu/~jurafsky/slp3/ **Core**

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016).

Data Mining: Practical machine learning tools and techniques, 4th edition. Morgan Kaufmann. See: https://ml.cms.waikato.ac.nz/index.html

PLUS: research conference papers, websites

Assessments

UNIVERSITY OF LEEDS

Assessment 1: online test 1 (30%), 1 hr, week 5

Assessment 2: individual report (70%), week 10

For the Report, you will develop a research project proposal, using data mining and text analytics theory, methods and technologies for a practical application of your choice.

See EPSRC guidance on writing research project proposals

AI research project proposal

UNIVERSITY OF LEEDS

SEE Minerva Assessment Overview

Proposed research and its context to include:

- Research hypothesis & objectives,
- Background,
- Contribution to knowledge,
- Programme and methodology
- Pilot Study

and **Workplan** diagram, eg Gantt Chart [additional 1 page]

APPENDIX: your use of DM+TA in writing the proposal

Programme and methodology

UNIVERSITY OF LEEDS

The research work programme should make use of an appropriate methodology for AI projects, such as CRISP-DM; and should include use of at least two data mining and/or text analytics methods, techniques or resources introduced in this module.

Data Mining / Machine Learning

UNIVERSITY OF LEEDS

Machine Learning: focus on ML algorithms, optimal accuracy

Data Mining: applied ML, with a focus on:

- ML as part of a toolkit to tackle practical problems
- Data collection, understanding, annotation, “wrangling”
- “Data analysts typically spend the majority of their time in the process of data wrangling compared to the actual analysis of the data” https://en.wikipedia.org/wiki/Data_wrangling
- CRISP-DM – “modelling” (ML) is only 1 of 6 phases

CRISP-DM: 6 Phases

UNIVERSITY OF LEEDS

Business Understanding

- Understanding project objectives and requirements
- Data mining problem definition

Data Understanding

- Initial data collection and familiarization
- Identify data quality issues
- Initial, obvious results

Data Preparation

- Record and attribute selection
- Data cleansing

Modeling

- Run the data analysis and data mining tools

Evaluation

- Determine if results meet business objectives
- Identify business issues that should have been addressed earlier

Deployment

- Put the resulting models into practice
- Set up for repeated/continuous mining of the data

Text Analytics

UNIVERSITY OF LEEDS

Data Mining applied to text … aka Text Mining, or …

Computational Linguistics / Natural Language Processing / Speech and Language Processing / Corpus Linguistics

- CL/NLP: focus on theory, algorithms
- TA: CL/NLP as part of a toolkit to tackle practical problems,

and text data collection, understanding, annotation, wrangling

Text data (CORPUS) is mapped to number vectors for ML (embeddings)

The Language Machine

UNIVERSITY OF LEEDS

"This book, commissioned by The British Council from Eric Atwell at the University of Leeds, explores some of the technological, social and educational implications of language machines in the years to come. …

This book provides a survey of the current state of speech and language technology … highlighting the histories and academic disciplines contributing to their development; it examines the components and technologies; possible pitfalls; main developers; current and potential uses; predicted developments; and paints some likely scenarios for the future impact of the language machine."

25 years old, but theoretical concepts are still relevant …

Linguistics: science of language

UNIVERSITY OF LEEDS

Phonetics: the study of speech production, perception, and analysis from an acoustic and a physiological point of view.

Lexis: the study of words or vocabulary items in a language, with individual meaning and grammatical function.

Syntax: the study of the grammatical arrangement of words and morphemes in the sentences of a language or languages.

Semantics: the study of meaning in language, the relationship between words and sentences and their meanings.

Pragmatics: the analysis of language in practice, taking account of the context of language use.

Discourse: the analysis of linguistic phenomena that range over more than one utterance in a discourse or dialogue

Why develop text analytics?

UNIVERSITY OF LEEDS

- Computer models of language
- Computerised language resources: corpus, dictionary,…
- Natural communication between people and computers
- Assisting communication between people: MT, social media
- Wealth creation: Government and Industry interest

Challenges for text analytics

UNIVERSITY OF LEEDS

- expensive to compete: Google, Apple, Amazon, Microsoft
- difficult to elicit user requirements: users don’t know
- high customer expectations: “natural English language”
- not appropriate for some tasks, eg spreadsheets?
- we need to rethink how we approach i/o, eg keyboards?
- we need training and time to learn to use new methods
- many applications involve all of the above

UK and EU research in text analytics

UNIVERSITY OF LEEDS

- UK: Engineering and Physical Science Research Council eg

  NLP working together with Arabic and Islamic Studies

  Making Sense: Detecting Terrorist Activities

—

- EU funds research projects with several partners, eg

  EduBots: chatbots in HE – 4 unis (Leeds ++), 2 companies

- International research agencies, eg Dubai Future Foundation:

KAMAL Health: Knowledge-Augmented Multi-Modal Arabic LLMs for Healthcare

The BT Technology Calendar

UNIVERSITY OF LEEDS

- 2000: visual computer personalities on screens
- 2003 IT literacy essential for any employment
- 2005 full voice interaction with machines
- 2007 domestic robots; small, attractive
- 2012 robots for almost any job in home or hospital
- 2018 AI imitating thinking processes of the brain
- 2025 thought recognition i/o, human learning superseded
- 2030 human brain intelligence enhancement by link to AI

Examples of real applications

UNIVERSITY OF LEEDS

- “Soldiers in Bosnia ... wear a small computer on their chests and say to it “Hands up” or “Get out of the car” or other things that soldiers have cause to order Bosnian civilians to do.”

- “Text editing: ‘smart tools’ to check grammar, idioms, and style are now options available in many word processors.”

- “AltaVista, owned by computer giant Digital, launched a free machine translation service on the Internet”

- ‘Lufthansa has ALF, a friendly flight information service which holds conversations with callers at some 300 airports”

- “car and lorry drivers use voice commands to activate normal telephone services but also to get e-mail messages converted to listen to them on the move, to dictate replies …”

Summary

UNIVERSITY OF LEEDS

In this lecture you were introduced to

- lecturer: Eric Atwell http://www.comp.leeds.ac.uk/eric
- Overview of Data Mining and Text Analytics module
- Assessment: test, DMTA research project proposal
- Practical applications of data mning and text analytics

Texts you should read in week 1:

- E Atwell. 1999. The language machine. British Council.
- Assessment specification