COMP 3225/6253

Natural Language Processing

Introduction

Stuart E. Middleton

sem03@soton.ac.uk

University of Southampton

Overview

- Welcome
- What is Natural Language Processing?
- Module teaching approach - flipped learning
- Learning outcomes
- Module overview

- <break - discussion point>

- History of Natural Language Processing
- Demonstration - Neural Machine Translation
- Reading for this lecture

Welcome

- Your NLP module lecturer team
  - Dr Stuart Middleton <sem03@soton.ac.uk> (module leader)
  - Dr Mercedes Arguello Casteleiro <mac1r22@soton.ac.uk>

- Dr Stuart Middleton
  - Associate Professor, AIC Research Group
  - Research work focusses on Natural Language Processing applied to areas including Cybercrime, Defence, Mental Health and Environmental Science

- Dr Mercedes Arguello Casteleiro
  - Lecturer, Digital Health and Biomedical Engineering (DHBE) Research Group
  - Research work focusses on Semantic Web and Deep Learning for Natural Language Processing

What is Natural Language Processing?

- Association for Computational Linguistics (ACL) definition
  - Computational Linguistics, or Natural Language Processing (NLP), is the scientific study of language from a computational perspective
- Some applications of NLP for textual data

• Spelling and Grammar Checking	MS Word
• Machine Translation	Google Translate
• Question Answering	Google Search, IBM Watson
• Conversational Agents / Dialogue Systems	Amazon Echo & Alexa
• Natural Language Generation	BBC 2019 Election News
• Information Extraction	
• Text Summarization	
• Machine Reading Comprehension	
• ...	

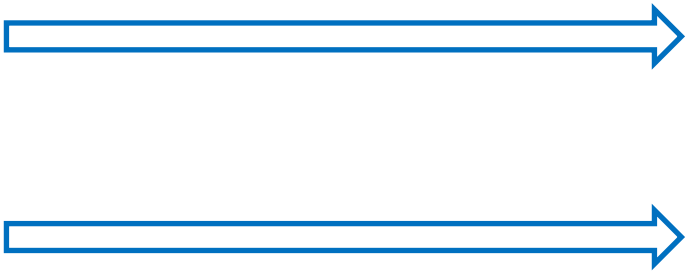
Module teaching approach - flipped learning

- Traditional learning
  - Go to lecture to get content; Understand it at home
- What is flipped learning?
  - Content delivered to students at home each week (video, reading)
  - Students watch & read weekly content **BEFORE** interactive sessions
  - Interactive session with lecturers help students understand the content from previous weeks

- How will it work for this module?

Weekly directed reading
(chapter/pages on module wiki)
average 3 - 4 hours a week

2 x Lecture videos
(Panopto links on module wiki)
average 1.5 hours a week



Interactive sessions
1 x Activity + 1 x Q&A
average 2 hours a week

We expect last weeks content to be watched before interactive sessions

Activity and Q&A lecture sessions face to face bring a laptop

Module teaching approach - flipped learning

- Traditional learning
  - Go to lecture to get content; Understand it at home
- What is flipped learning?
  - Content delivered to students at home each week (video, reading)
  - Students watch & read weekly content **BEFORE** interactive sessions
  - Interactive session with lecturers help students understand the content from previous weeks

- How will it work for this module?

Run at home labs (3 NLP labs + python lab)
(jupyter notebook tutorial via module wiki)
8 hours total home lab work not inc. model training time

Practice your practical NLP skills ready for coursework assignment

python lab >> booster to students rusty on python
NLP labs >> core content needed for coursework

Optional weekly drop-in lab support sessions run by demonstrators
book in via MS Forms
(20 student limit each week)

Learning outcomes

- At the end of this module you should be able to ...

- Theory
  - Show an understanding of concepts, tools and approaches for textual data
  - Have knowledge of the underlying algorithmic and linguistic basis for NLP
  - Show an understand of common NLP algorithms
  - Describe and discuss different subareas of NLP
  - Understand the potential and limitations of NLP in application areas

- Practice [UG]
  - Process text corpora ready for NLP
  - Implement NLP algorithms and techniques

- Practice [MSc]
  - Describe and critically appraise the different subareas of NLP
  - Show an appreciation of the landscape of tools used for NLP within application areas

Module overview

- Theory lectures
  - Reading (3 - 4 hours per week)
  - Pre-recorded lecture content (1.5 hours per week)
  - Interactive sessions (2 hours per week)
    - Practical activity for last weeks content [laptop required]
    - Q&A focusing on content up to and including last week
  - Assessment: Exam 75%
- Practical labs
  - Tutorial style 'run at home' labs + drop-in sessions (8 hours total)
  - Coursework (45 hours)
  - Assessment: Coursework 25%
- Revision
  - Revision (10 hours)

Module overview

- Introduction [Stuart]
  - This lecture!
- Working with Text Corpora [Stuart, Mercedes]
  - Words; Regular Expressions; Evaluation and Linguistic Resources
- Language Modelling and Parts of Speech Tagging [Stuart, Mercedes]
  - Language modelling; Parts of Speech Tagging, Named Entity Recognition
- Vector Semantics and Embeddings [Stuart]
  - Lexical and Vector Semantics; TF-IDF and Word2Vec
- Coursework Overview [Stuart]
  - 2 x regex tasks
  - 1 x named entity recognition task
  - 1 x named entity recognition + regex task
  - Python 3 code submission, automatic marking, 20 submission attempts allowed

Module overview

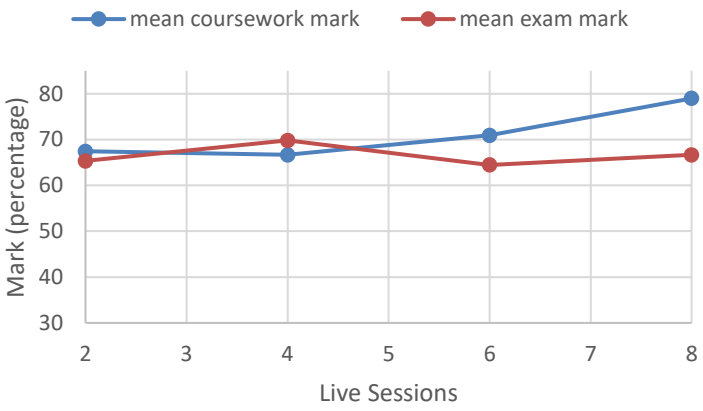
- Sequence Processing with Recurrent Neural Networks [Stuart]
  - RNN; Sequence Processing, Transformer and Attention; Machine Translation and Encoder-Decoder models
- Syntactic and Semantic Parsing [Stuart]
  - Constituency Grammars; Syntactic parsing and text chunking; Dependency Parsing; Word Senses and WordNet
- Information Extraction [Stuart]
  - Relation Extraction; Temporal, Event and Location Extraction
- Applications of NLP [Stuart]
  - Machine Translation; Semantic Role Labelling; Question Answering
- Revision [Stuart]

Module overview

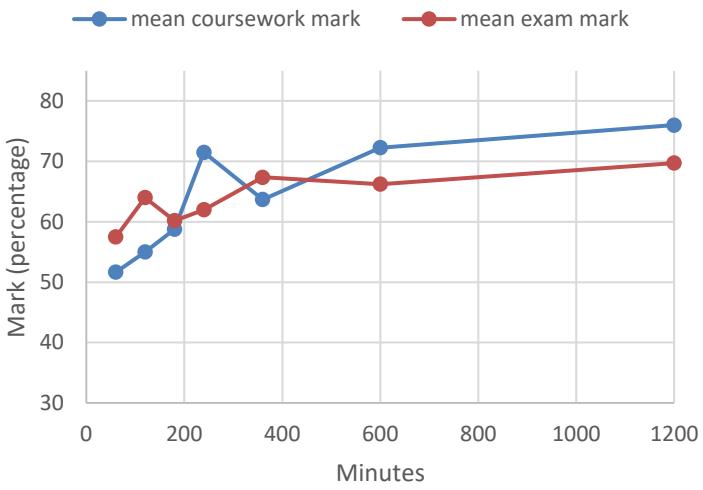
- Labs
  - All labs written using Python 3, Scikit Learn and Tensorflow, running via Jupyter Notebook locally on your laptops
  - Lab 1 Python Basics [Radu]
  - Lab 2 Regex [Stuart]
  - Lab 3 Named Entity Recognition using CRF [Stuart]
  - Lab 4 Neural Machine Translation [Stuart]
- Drop-in lab demonstrator led sessions
  - Book via MS Form booking for (20 students max a week)
    - see Module wiki for link
  - Python programming help >> Students with rusty Python skills
  - NLP labs >> Core material useful for coursework
  - Work on the labs at home, drop-in if you need further support

Analysis - Engagement VS Marks [2021]

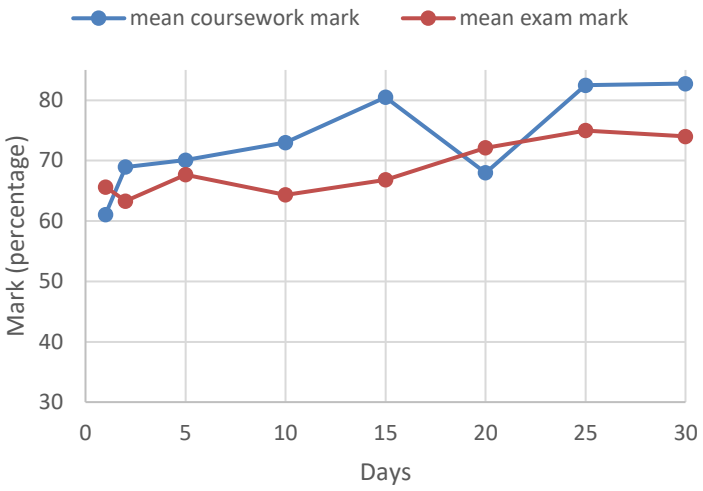
Interactive sessions attended live (sessions)



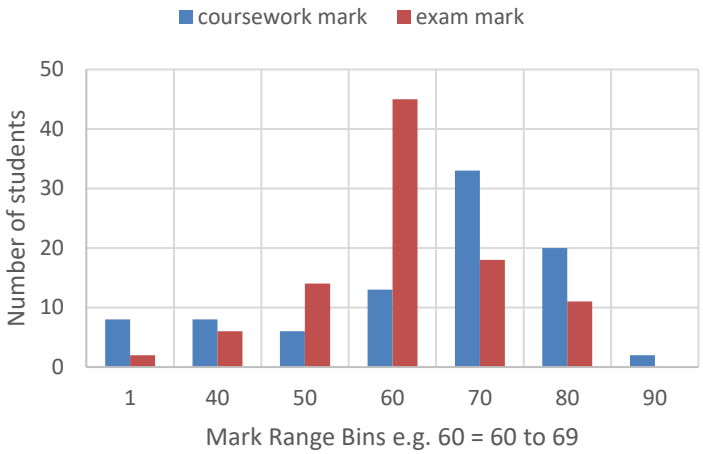
Panopto videos watched (minutes)



First coursework submission attempt (days before final attempt)

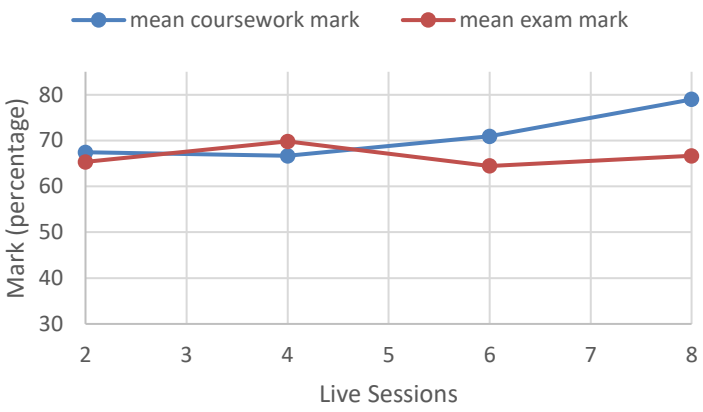


Mark Distribution

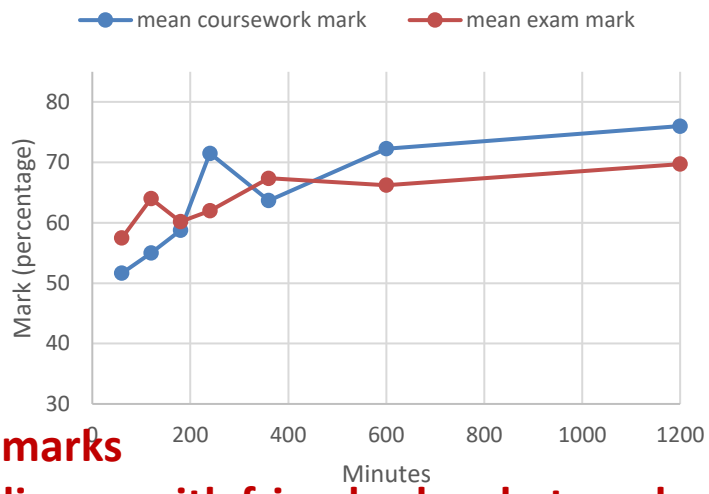


Analysis - Engagement VS Marks [2021]

Interactive sessions attended live (sessions)



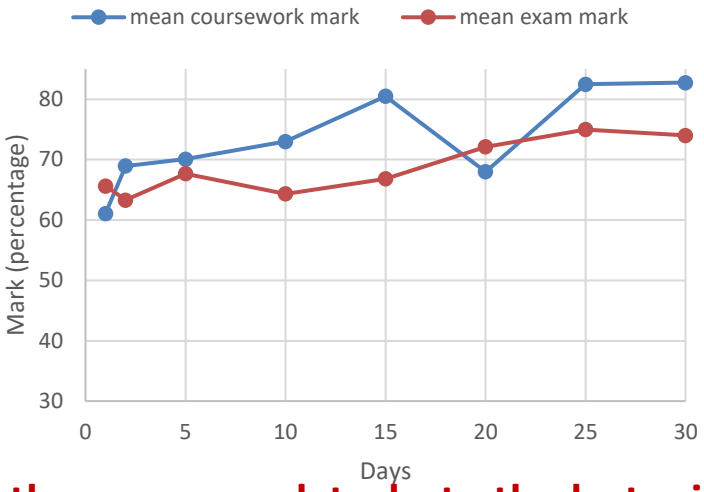
Panopto videos watched (minutes)



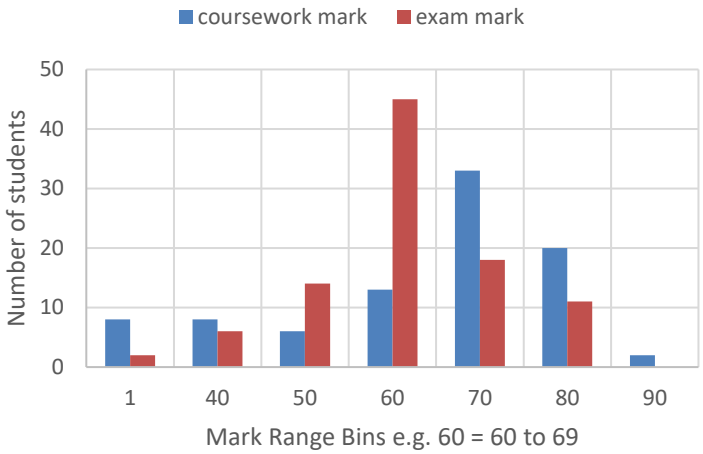
Regular engagement is key to help you get top marks

Plan it - schedule quiet reading time, watch & discuss with friends, do what works for you

First coursework submission attempt (days before final attempt)



Mark Distribution



Leaving the coursework tasks to the last minute correlates to low marks (!)

Break

- Panopto Quiz - discussion point

- Which of the following applications of NLP today learn to understand the text they are processing?

Machine Translation
Question Answering
Spelling and Grammar Checking
Conversational Agents / Dialogue Systems
Natural Language Generation
Information Extraction
Text Summarization

Break

- Panopto Quiz - discussion point

- Which of the following applications of NLP today learn to understand the text they are processing?

Machine Translation >> Map English text X to Russian text Y >> **No**
Question Answering >> Map questions X to answer Y >> **No**
Spelling and Grammar Checking >> Lookup word X to suggest Y >> **No**
Conversational Agents / Dialogue Systems >> Map chat X + context Y to response Z >> **No**
Natural Language Generation >> Map data X to Story Y >> **No**
Information Extraction >> Given document X extract knowledge base Y >> **No**
Text Summarization >> Given document X create shorter summary Y >> **No**

Today's NLP algorithms are really good at learning the right response to a text stimuli.

Human level machine understanding is still science fiction though.

Can bigger and deeper learning get us there? Only time will tell ...

History of Natural Language Processing

"He who controls the present, controls the past. He who controls the past, controls the future", George Orwell, 1984

- Foundations of NLP - 1940s and 1950s
  - Automaton >> Context Free Grammar
  - Probabilistic models >> Motivated by entropy in thermodynamics
- Symbolic or Stochastic - 1950s to 1960s
  - Symbolic parsers >> Reasoning and Logic
  - Stochastic >> Bayesian-based language likelihood predictions
- Symbolic or Stochastic - 1970s to 1980s
  - Stochastic >> Hidden Markov Models
  - Logic >> Functional Grammars
  - Natural language understanding >> Parsing + Semantics
  - Discourse modelling >> Speech Acts + Discourse

History of Natural Language Processing

- Empiricism and Finite State Models - 1980s and 1990s
  - Return to probabilistic and data driven methods
  - Enablers: Speed and memory growth in computers; Web corpora
- Machine learning - 2000s
  - Statistical approaches to machine translation and topic modelling
  - Enablers: Annotated corpora such as LDC (Treebank, Propbank etc.)
- Deep learning - 2010s
  - Multi-layer deep neural models
  - Enablers: Unlabelled or weakly annotated web scale corpora
  - Enablers: Advances in deep learning from image processing; GPU clusters

Demonstration - Neural Machine Translation

- Demonstration
  - Russian to English machine translation of online cybercrime posts
- Source posts
  - Posts crawled from Russian cybercrime forums
- Experiment
  - Professional translator (Human)
  - Commercial service (Google Translate)
  - Deep-learning model (Transformer) trained on Wikipedia bitext
  - Deep-learning model (Transformer) trained on Wikipedia + cyber bitext
- Computing hardware used to train and run models
  - UoS IRIDIS GPU cluster (node with 2xGPU card)

https://southampton.cloud.panopto.eu/Panopto/Pages/Viewer.aspx?id=37bccdd1-5c9f-46b9-bebb-acae00fb125b

Required Reading

- Introduction and history of NLP [optional]
  - Jurafsky and Martin, Speech and Language Processing, 2nd edition, Prentice Hall, 2009 >> Chapter 1

Questions

- Panopto Quiz - 1 minute brainstorm for interactive questions

  Please write down in Panopto quiz in **1 minute** two or three questions that you would like to have answered at the next interactive session.

  Do it **right now** while its fresh.

  Take a screen shot of your questions and **bring them with you** at the interactive session so you have something to ask.