Project choices

Leon Derczynski

Innopolis University

Project

- 40% of the overall mark
- You’re welcome to work in groups of 1-3
  - Bigger group means you need a better assignment

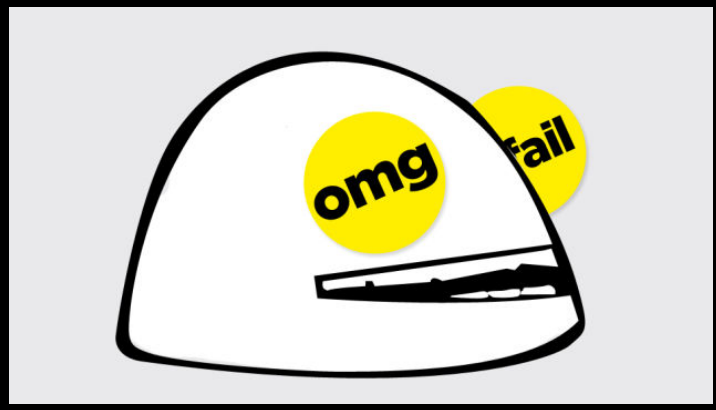
NER with gazetteers

- Summary
  - Named entity recognition, adding lists of some terms according to their type (cities, footballers, biscuits)

- Challenges
  - How to add gazetteer knowledge to normal features?
  - Feature sparsity could be an issue
- Input
  - CoNLL training data with NEs, Gazetteer lists
- Output
  - Test data with tokens labelled with entity type

NER for social media

- Summary
  - Find NEs in social media
- Challenges
  - See week 3 :) Unusual terms, unreliable case, mis-spelling
- Input
  - CoNLL-format training data with NEs, gazetteer lists
- Output
  - Entities from social media



NER for a new entity type

- Summary
  - Pick an unusual entity type; build recogniser for them
    - Types of chocolate
    - Names of prescription medicine
    - Extremist political parties
- Challenges
  - No data
- Input
  - Data that you find and annotate yourself (half an hour's work can be more than enough)
- Output
  - Tool for identifying these entities automatically

Sentiment for reviews

- Summary
  - Is a review positive or negative?
- Challenges
  - Not all the information is in unigrams
  - Irony / sarcasm (Oh great, Jennifer Aniston again!)
- Input
  - Reviews (for something.. Film? Restaurant? Women's Pens?)
  - Labels
- Output
  - Positive / negative (or if you prefer, p / n / neutral)

Check reactions

- Formally called “Stance Detection”
- Support, deny, query or comment on a claim?
- LSTM classification could work (sample code for this from week 2)
- .. so could simpler methods: building lists of words that match each “stance”, then using these as features

- Data in “RumourEval” - Task A

Find rumours

- Modern problem: what news stories are real?
- RumourEval “Task B” http://alt.qcri.org/semeval2017/task8/
- Fake News Challenge http://www.fakenewschallenge.org/
- Basically a classification challenge
- Tough – needs world knowledge!

Author gender prediction

- Summary
  - What's the gender of the author?
- Challenges
  - Assumes that genders have styles
- Input
  - Text written by men, text written by women (or any set of genders G such that there is text for $g \in G$)
- Output
  - For an input text, a label g

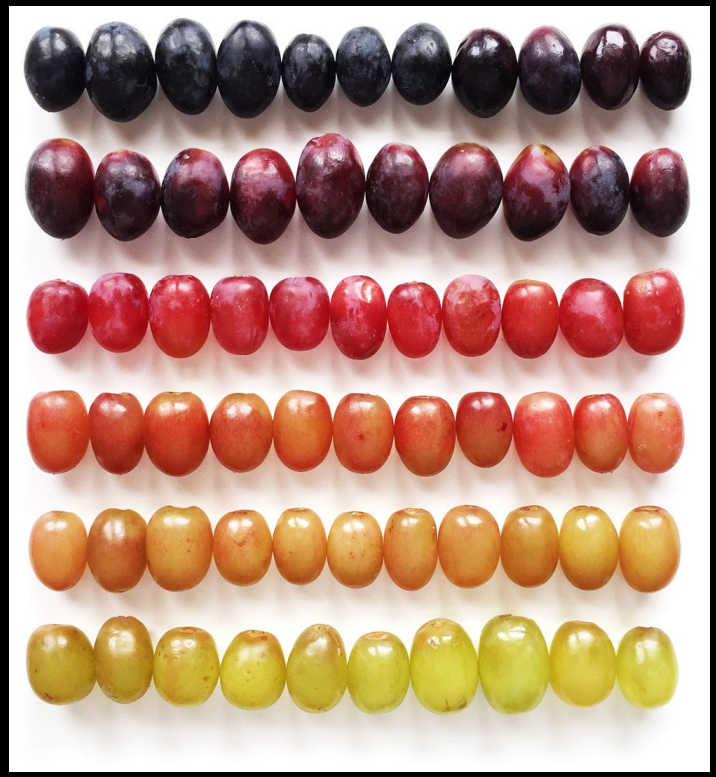
PoS tagger for a new language

- Summary
  - Here's a new language (Dothraki?). Let's PoS tag it!
- Challenges
  - No data! That's OK; state-of-the-art accuracy can be attained with ~2 hours annotation

    “Learning a Part-of-Speech Tagger from Two Hours of Annotation”, Garrette and Baldridge, NAACL 2013
  - ...We don't need state-of-the-art :)
- Input
  - Text you've annotated with POS (following e.g. universal scheme)
- Output
  - A totally new tool for handling an “unresourced” language

Mine new entity terms

- Summary
  - Find new entity names, given a few examples of one type
    - e.g. given Lokomotive Moscow, Amkar Perm, the system should return things like Dinamo Moscow
  - You can define this subjectively
    - Houseplants
    - Words with five letters
    - Things I would like for christmas
- Challenges
  - Slightly technical (LLDA has a good tutorial!)
- Input
  - Lots of unlabelled text
  - A few examples of entities you like
- Output
  - Untold cornucopia of good-to-mediocre examples of that entity



Generative Eliza

- Build a copy of Eliza in Python
- Find some dialogues as training data
  - I have a lot of conversation scripts, ask!
- Learn a language model
- Output Eliza-like sentences
- One idea:
  - Train a seeded NLG system, like in the LSTM language model tutorial, based on some other conversation scripts
  - Make it talk with Eliza
  - Record the responses, so you have many Eliza conversations
  - Use this output, as a training set for Eliza responses
  - The resulting model can generate Eliza's side of the dialogue

More ideas

- SemEval has some cool tasks!
  - Try something with existing data – can you beat the state of the art?
    - http://alt.qcri.org/semeval2017/index.php?id=tasks
    - http://alt.qcri.org/semeval2018/index.php?id=tasks

Project format

- Write as an academic paper
  - Use the LREC 2020 style files
  - https://lrec2020.lrec-conf.org/en/submission2020/authors-kit/
  - Results will be published informally
  - You’re welcome to submit to LREC with my help
- Submit a project proposal first
  - Due ASAP
  - This describes the problem you’d like to work on
  - I’ll make sure you approach the right-sized problem

Project format

- Main sections:
  - Introduction
  - Background (literature, similar previous work)
  - Method
    - Dataset
    - Baseline: a simple approach
    - Your NLP approach
  - Analysis
    - Performance scores, e.g. accuracy, F1
    - What worked, what didn’t work, and why
    - Did you have enough data?
    - What would you do differently next time
  - Conclusion
- It’s a good idea to include examples to help communicate the problem, and also a graph or two to describe performance.

Course wrap-up

- Project assignment due **November 18**
- Mail me for any corpora/annotations, there are *many*

- Input:
  - Code or a link to a colab
  - Documentation (4-page paper)
- Output:
  - Sweet, sweet ECTS

- Thanks for participating!