Syllabus NLP for social sciences (and other applications)

The aim of this group is to train participants on the most recent techniques of natural language processing, text analysis, scraping, network theory and machine learning on graphs while enabling them to implement these techniques on different applied projects.

The program is divided into two parts. A first part, more technical and common to all the participants, that deals with NLP techniques, scraping 101 and fundamentals of network theory + learning on graph structures. We will approach these concepts from courses and materials available online, and we will discuss them together in online meetings.
The second part will be devoted to the application of NLP techniques to topics of interests and it will be possible to form various groups on different topics. We plan to create at least one group that will apply NLP techniques to political texts and political communication in general. Other groups may be created as well.

We plan to meet once every two/three weeks. No previous knowledge of NLP and text analysis is required, even though a familiarity with Python might be preferable.

Tentative programme:

- Topics in NLP

The course we will mainly follow is “Deeplearning.ai’s NLP Specialization” on Coursera (https://www.coursera.org/specializations/natural-language-processing) + adding some more technical stuff .

Topics covered:

NLP with Classification & Vector Spaces

- Supervised ML & Sentiment Analysis
- Vocabulary & Feature Extraction
- Negative & Positive Frequencies
- Feature Extraction with Frequencies
- Preprocessing
- Logistic Regression (LR) Overview
- Training LR
- Testing LR
- Create a Cost Function for LR
- Gradient Descent with LR
- Bayes' Rule
- Naive Bayes
- Smoothing & Laplacian Smoothing
- Log Likelihood
- Training Naive Bayes
- Testing Naive Bayes
- Naive Bayes Assumptions
- Error Analysis
- Vector Space Models
- Word by Word & Word by Doc
- Euclidean Distance
- Cosine Similarity
- Manipulating Word Embeddings (King + Man - Queen = Woman)

- Visualization and PCA (What about nonlinear dimensionality reduction techniques, e.g. t-SNE, UMAP;
- Transforming Word vectors (Rotation matrices in R2)
- K-Nearest Neighbors (kNN)
- Hash Tables & Hash Functions
- Locality Sensitive Hashing (LSH)
- Approximate kNN
- Searching Documents
- Word Translation 101

NLP with Probabilistic Models

- Building an Autocorrect Model
- Minimum Edit Distance with Algorithms
- Part of Speech Tagging (POS)
- Markov Chains
- Hidden Markov Models
- Transition and Emission Matrices
- The Viterbi Algorithm (Initialization, Forward Pass, Backward Pass)
- N-Grams
- Sequence Probabilities (+ starting & ending a sentence)
- N-Gram Language Model
- Out of Vocabulary (OOV) Words

NLP with Sequence Models

- Neural Networks for Sentiment Analysis
- Trax Library Introduction: Neural Networks & Layers
- Dense and ReLU Layers
- Serial + Other Layer
- Traditional Language models
- Recurrent Neural Networks
- Applications of RNNs
- Math in Simple RNNs
- Cost Function for RNNs
- Gated Recurrent Units (GRUs)
- Deep and Bi-directional RNNs
- RNNs and Vanishing Gradients
- Introduction to LSTMs and their architecture
- Introduction to Named Entity Recognition (NER)
- Training NERs: Data Processing
- Siamese Networks & their architecture
- Cost Function
- Triplets (triplet loss..)
- Computing The Cost I
- Computing The Cost II
- One Shot Learning

NLP with Attention Models

- Seq2seq
- Alignment
- Attention

- Setup for Machine Translation
- Training an NMT with Attention
- Evaluation for Machine Translation
- Sampling and Decoding
- Transformers vs RNNs
- Transformer Applications
- Dot-Product Attention
- Causal Attention
- Multi-head Attention
- Transformer Decoder
- Transfer Learning in NLP
- ELMo, GPT, BERT, T5
- Bidirectional Encoder Representations from Transformers (BERT)
- BERT Objective
- Fine tuning BERT
- Transformer: T5
- Multi-Task Training Strategy
- GLUE Benchmark
- Question Answering
- Tasks with Long Sequences
- Transformer Complexity
- LSH Attention
- Motivation for Reversible Layers: Memory!
- Reversible Residual Layers
- Reformer

- Topics in ML on Graphs (and Intro on Network Theory)

The course we will mainly follow is Stanford's "Machine Learning with Graphs" class by Jure Leskovec (http://web.stanford.edu/class/cs224w/).

Also, an awesome book by William L. Hamilton :
https://www.cs.mcgill.ca/~wlh/grl_book/files/GRL_Book.pdf

Topics covered:

Network Theory Fundamentals

- Structure of Graphs
- Properties of Networks and Random Graph Models
- Snap.py & Google Cloud basics
- Motifs and Structural Roles in Networks
- Community Structure in Networks
- Spectral Clustering
- Page Rank algorithm
- Network effects and Cascading Behaviour
- Probabilistic Contagion and Models of Influence
- Influence Maximization in Networks
- Outbreak Detection in Networks
- Network Evolution

Learning on Graphs

- Message Passing and Node Classification
- Graph Representation Learning
- Graph Neural Networks
- Deep Generative Models for Graphs
- Reasoning over Knowledge Graphs
- Limitations of Graph Neural Networks
- Applications of Graph Neural Networks

- Scraping

Material we plan to use:

Web Scraping in Python
https://www.datacamp.com/courses/web-scraping-with-python Datacamp

https://docs.scrapy.org/en/latest/intro/tutorial.html Scrapy Tutorials

Twint Project
https://github.com/twintproject/twint

Fbcrawl Project
https://github.com/rugantio/fbcrawl (there might be the chance to contribute actively to this repo, the owner is a former MLJC member)

- Applications:

- Application to social sciences and political texts

References:

Gentzkow, M., Kelly, B., & Taddy, M. (2019). **Text as data.** *Journal of Economic Literature*, *57*(3), 535-74. (link here: https://web.stanford.edu/~gentzkow/research/text-as-data.pdf)

Grimmer, J., & Stewart, B. M. (2013). **Text as data: The promise and pitfalls of automatic content analysis methods for political texts.** *Political analysis*, *21*(3), 267-297.

Gentzkow, M., Shapiro, J. M., & Taddy, M. (2019**). Measuring group differences in high-dimensional choices: method and application to congressional speech.** *Econometrica*, *87*(4), 1307-1340

Le Pennec, C. (2020**). *Strategic Campaign Communication: Evidence from 30,000 Candidate Manifestos*** (No. 2020-05). Monash University, SoDa Laboratories.

Iyyer, M., Enns, P., Boyd-Graber, J., & Resnik, P. (2014, June). Political ideology detection using recursive neural networks.

Ceron, A., Curini, L., Iacus, S. M., & Porro, G. (2014). Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New media & society*, *16*(2), 340-358.

Liguori, M., Steccolini, I., & Rota, S. (2018). Studying administrative reforms through textual analysis: the case of Italian central government accounting. *International Review of Administrative Sciences*, *84*(2), 308-333.

Field, A., Kliger, D., Wintner, S., Pan, J., Jurafsky, D., & Tsvetkov, Y. (2018). Framing and agenda-setting in russian news: a computational analysis of intricate political strategies. *arXiv preprint arXiv:1808.09386*. link: https://arxiv.org/pdf/1808.09386

- Application: prediction of mental health from social media

References:

Chancellor, S., & De Choudhury, M. (2020). Methods in predictive techniques for mental health status on social media: a critical review.

Coppersmith, G., Dredze, M., & Harman, C. (2014, June). Quantifying mental health signals in Twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality* (pp. 51-60).

- Application: Fake news detection on social media + analysis of echo chambers with focus on politics (former Social Resistance MLJC Project)

Previous Materials
https://docs.google.com/document/d/1IzQ4NrrKEY232ZjeYKxGDtaYPupQHa4leeFTQTYnjd0/edit

References:

non-technical (although nice introduction in ITA)
Blog Micromega, Oltre le echo chamber, una bussola filosofica per il mare della complessità
http://lameladinewton-micromega.blogautore.espresso.repubblica.it/2018/12/10/oltre-le-echo-chamber-una-bussola-filosofica-per-il-mare-della-complessita/

Quattrociocchi W. et al, Echo Chambers on Facebook
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2795110

A.Acerbi et al., Cognitive attraction and online misinformation https://www.nature.com/articles/s41599-019-0224-y

X.Qiu et al, Limited individual attention and online virality of low-quality information
https://www.nature.com/articles/s41562-017-0132

S. Vosoughi et al, The spread of true and false news online
https://science.sciencemag.org/content/359/6380/1146

-- What Twitter does (probably an improved version!) wrt to fake news ---
Bronstein et al, Fake News Detection on Social Media using Geometric Deep Learning
https://arxiv.org/abs/1902.06673