DSBA M2 Assignment

Part I: Network analysis

Introduction

- In th first Part 2, you will replicate a well known network analysis, with different data and some twists.
- Data: The data is to be found at: https://github.com/SDS-AAU/SDS-master/tree/master/00_data/network_krackhard (Hint: You neet to download the raw data)

Data: What do I get?

Background

Let the fun begin. You will analyze network datacollected from the managers of a high-tec company. This dataset, originating from the paper below, is widely used in research on organizational networks. Time to give it a shot as well. Krackhardt D. (1987). Cognitive social structures. Social Networks, 9, 104-134. The company manufactured high-tech equipment on the west coast of the United States and had just over 100 employees with 21 managers. Each manager was asked to whom do you go to for advice and who is your friend, to whom do you report was taken from company documents. Description

The dataset includes 4 files - 3xKrack-High-Tec and 1x High-Tec-Attributes. Krack-High-Tec includes the following three 21x3 text matrices:

- ADVICE, directed, binary
- FRIENDSHIP, directed, binary
- REPORTS_TO, directed, binary

Column 1 contains the ID of the ego (from where the edge starts), and column 2 the alter (to which the edge goes). Column 3 indicates the presence (=1) or absence (=0) of an edge.

High-Tec-Attributes includes one 21x4 valued matrix.

- ID: Numeric ID of the manager
- AGE: The managers age (in years)
- TENURE: The length of service or tenure (in years)
- LEVEL: The level in the corporate hierarchy (coded 1,2 and 3; 1 = CEO, 2 = Vice President, 3 = manager)
- DEPT: The department (coded 1,2,3,4 with the CEO in department 0, ie not in a department)

Tasks

1. Create a network

- Generate network objects for the companies organizational structure (reports to), friendship, advice
- This networks are generated from the corresponding edgelists
- Also attach node characteristics from the corresponding nodelist

2. Analysis

Make a little analysis on:

A: Network level characteristics. Find the overal network level of:

- Density
- Transistivity (Clustering Coefficient)
- Reciprocity

. . . for the different networks. Describe and interpret the results. Answer the following questions:

- Are relationships like friendship and advice giving usually reciprocal?
- Are friends of your friends also your friends?
- Are the employees generally more likely to be in a friendship or advice-seeking relationship?

B: Node level characteristics: Likewise, find out:

- Who is most popular in the networks. Who is the most wanted friend, and advice giver?
- Are managers in higher hirarchy more popular as friend, and advice giver?

C: Relational Characteristics: Answer the following questions:

- Are managers from the same 1. department, or on the same 2. hirarchy, 3. age, or 4. tenuere more likely to become friends or give advice? (hint: assortiativity related)
- Are friends more likely to give each others advice?

3. Visualization

Everything goes. Show us some pretty and informative plots. Choose what to plot, and how, on your own. Interpret the results and share some insights.

Part 2: NLP: Hate-speech and offensive language on Twitter

Figure 1: hatespeech

This assignment is less structured than previous individual assignments.

You are given a collection of approximately 25k tweets that have been manually (human) annotated. `class` denotes: 0 - hate speech, 1 - offensive language, 2 - neither

`https://github.com/SDS-AAU/SDS-2020/raw/master/M2/assignments/data/twitter_hate.zip`

1. Preprocessing and vectorizaion.

Justify your choices and explain possible alternatives (e.g. removing stopwords, identifying bi/tri-grams, removing verbs or use of stemming, lemmatization etc.) - Create a bag-of-words representation, apply TF-IDF and dimensionality reduction (LSA-topic modelling alternatively simply PCA or SVD) to transform your corpus into a feature matrix.

2. Explore and compare the 2 "classes of interest" - hate speech vs offensive language.

- Can you see differences by using simple count-based approaches?
- Can you identify themes (aka clusters / topics) that are specific for one class or another? Explore them using, e.g. simple crosstabs - topic vs. class and to get more detailed insights within-cluster top (TF-IDF) terms. (This step requires preprocessed/tokenized inputs).

3. Build an ML model that can predict hate speech

Use the ML pipeline (learned in M1) to build a classification model that can identify offensive language and hate speech. It is not an easy task to get good results. Experiment with different models on the two types of text-representations that you create in 2. Bonus: Explore missclassified hate speech tweets vs those correctly predicted. Can you find specific patterns? Can you observe some topics that are more prevalent in those that the model identifies correcly? The best-reported results for this dataset are.

Class	Precision
0	0.61
1	0.91
2	0.95
Overall	0.91

Here advanced NLP feature engineering has been used, and thus everything around an overall accuracy of 85 is fine. You will see that it is not easy to lift class 0 accuracy over 0.5 Good Luck!

Deliverables

Please submit a PDF or HTML version of your notebook on peergrade.io (if you submit HTML, please zip it before - large embedded HTMLs from cause crashing when oppened directly in peergrade). In adittion, feel free to include a colab link (non-mandatory)). Please make sure it runs without errors and others can access it (i.e. own test in "anonymous" setting in your browser).

This notebook should:

- It should solve the questions in an straightforward and elegant way.
- It should contain enough explanations to enable your fellow students (or others on a similar level of knowledge) to clearly understand what you are doing, why, what is the outcome, how to interpret it, and how to reconstruct the exercise. Be specific and understandable, but brief.

Further process and dates

- Hand in deadline: **02.11.2020, 11.55pm**

- You will receive an upload link on peergrade.io with concrete instructions.
- After the upload deadline, you will recieve an invitation to peergrade your fellows' exams on peergrade.io. You will be asked for the evaluation of 3 peer-assignments.