Graph-Based Analysis of Drama Character Interaction

T. Zhordaniya, M. Podryadchikova

Outline

1. Dataset: format and description

2. Hypothesis: some features of the character and her representation in the graph can be connected

- The subject of analysis: classes and variables

- Statistical tests

- SVM prediction

3. Conclusion

RusDraCor: format

Play = **graph**

Character = **node**

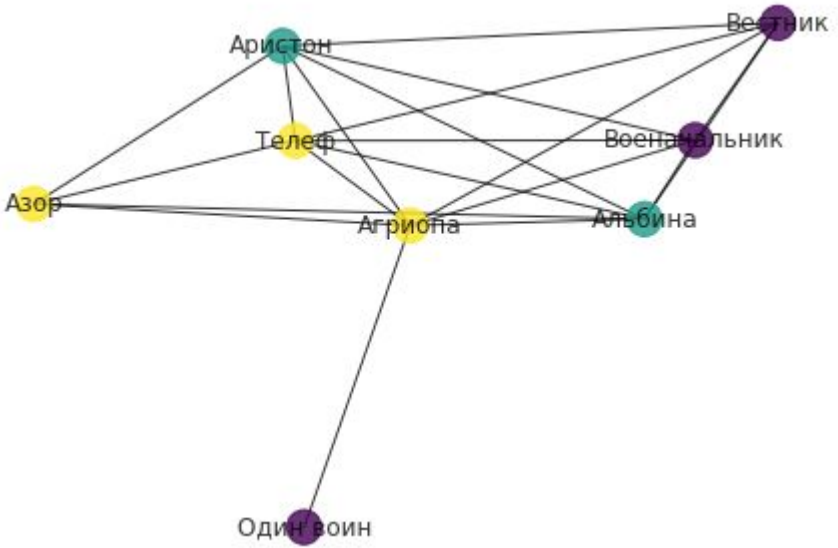
Co-occurence of characters in the play's scene = **edge** between these two characters

attributes:
- *gender* (M/F/other), *label* (character's name), *number of spoken words* — available in the corpus

- *importance* (lead/secondary/episode) — marked-up by us

N. Gogol, *Marriage*

V. Majkov, Agriopa (color ~ importance)



Main question

Is it possible to predict character's characteristics using the information about the corresponding node?

We decided to analyze how *gender/importance* of characters are connected with graph indicators.

What's the point?

"There are so many ~~books~~ plays. There is so little time"

distant reading — an approach in literary studies that applies computational methods to literary data, usually derived from large digital libraries, for the purposes of literary history and theory.

The subject of analysis

Gender: 232 Male, 102 Female characters for 50 plays.

Importance: 123 Main, 135 Supporting, 75 Episodic

Variables:

- betweenness centrality
- degree centrality
- PageRank
- percent of words said in the play

Hypotheses

for gender:

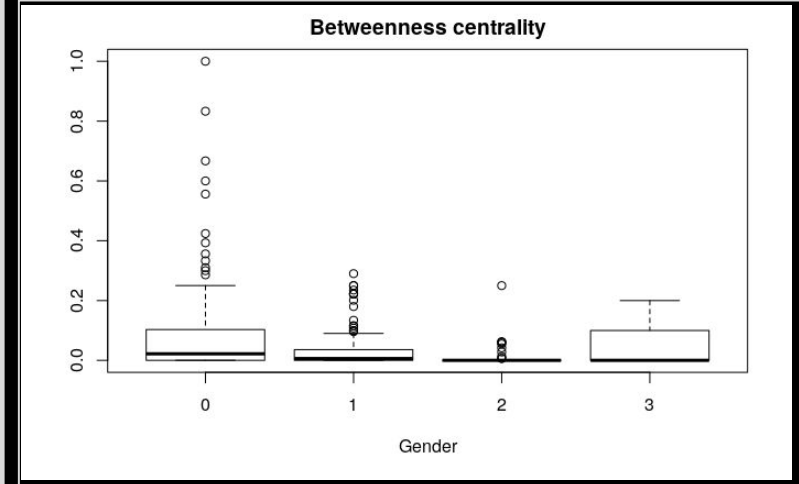
H0: There is no statistically significant difference between */variable_name/* values for male and female characters

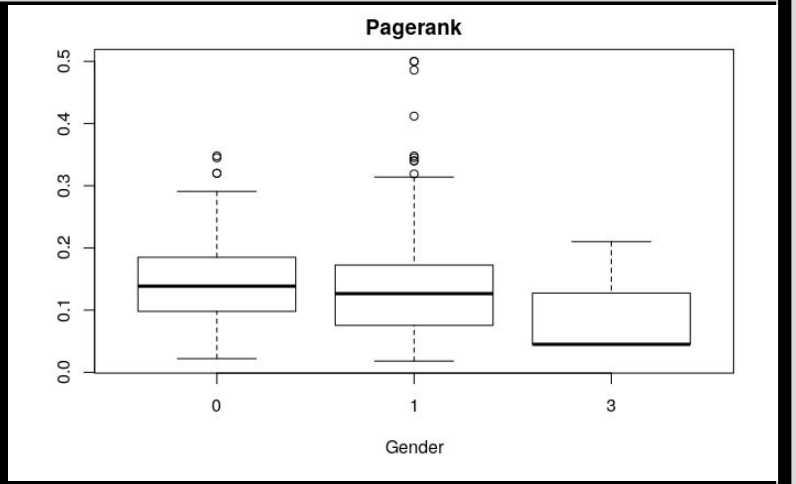
H1: There is statistically significant difference

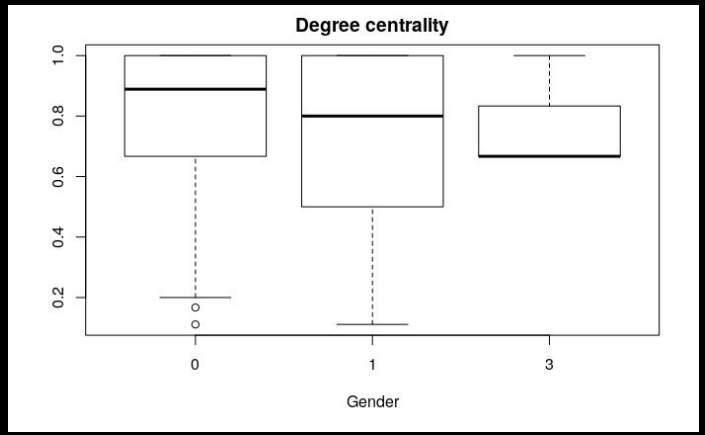
for importance:

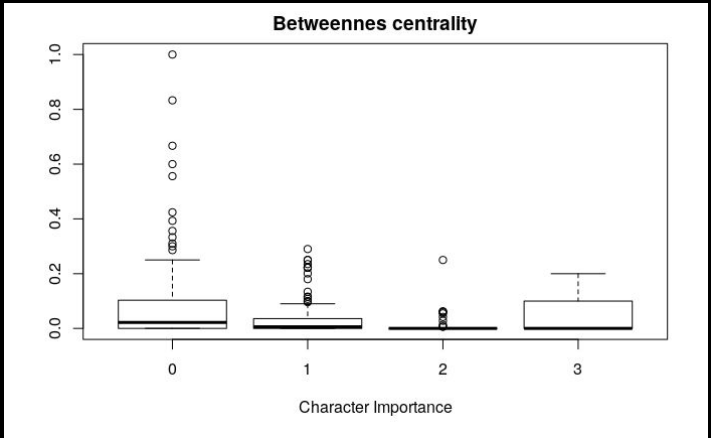
H0: There is no statistically significant difference between */variable_name/* values for all three levels of importance

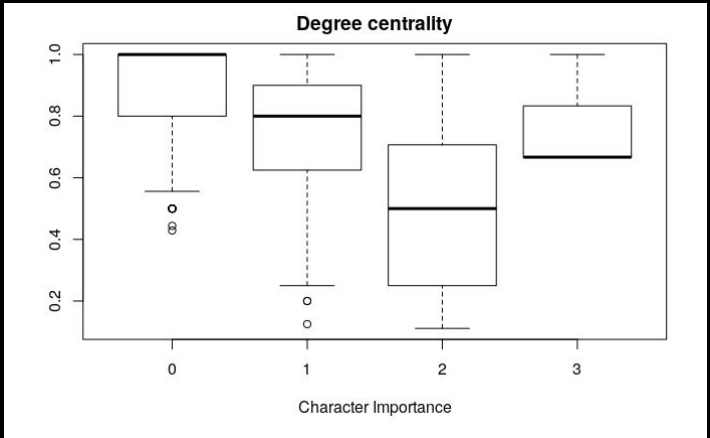
H1: At least two levels are different

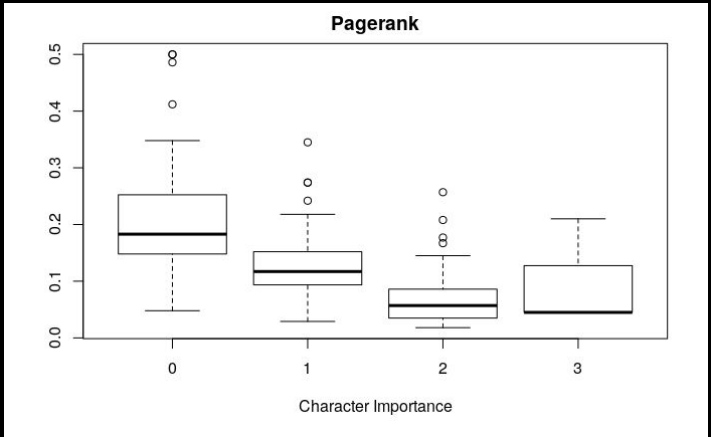

Betweenness centrality
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
Gender

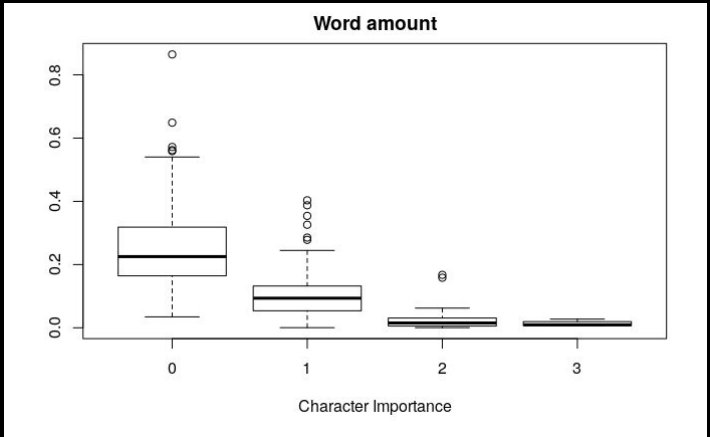

Pagerank
0.0
0.1
0.2
0.3
0.4
0.5
0
1
3
Gender


Degree centrality
0.2
0.4
0.6
0.8
1.0
0
1
3
Gender


Betweennes centrality
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
Character Importance


Degree centrality
0.2
0.4
0.6
0.8
1.0
0
1
2
3
Character Importance


Pagerank
0.0
0.1
0.2
0.3
0.4
0.5
0
1
2
3
Character Importance


Word amount
0.0
0.2
0.4
0.6
0.8
0
1
2
3
Character Importance

Results of statistical tests: Gender

For male and female characters, we cannot say that there is statistically significant difference (H0 is not rejected) for any of variables.

Results of statistical tests: Importance

For characters of different importance:

1. *Analysis of variance* proved that there is difference for all of variables.

2. According to *Tukey's range test*, **there is statistically significant difference** between all pairs for all variables except for betweenness centrality for pair "episodic-supportive".

Support Vector Machine model

Features for every character:

- degree centrality
- pagerank
- relative spoken word count

Target:

- importance class (lead, supporting, episode)

Model:

- *SVM* with ‘ovo’ decision function

Prediction results

- accuracy score = 0.70
- macro f1 score = 0.69
- micro f1 score = 0.70
- recall score = 0.70

Conclusion

As a result, we can see that we can distinguish main characters from episodic using graph methods.

We also did not discovered statistically significant differences between representation of men and women as nodes — even though there are less female characters, they are often important for the play.

What can be improved

- Markup quality (!)
- More data
- Less obvious hypothesis