ISMT-117 Final Project
Paige Lee
8/7/20

Characterizing the writing styles of male and female Russian authors

In this project, I aim to apply the techniques learned in this course to the classification and characterization of Russian texts on the basis of the author's gender. The study of gender recognition based on texts has a number of applications from historical literary analysis to human-computer interaction. The question of whether an author's gender can be determined from their writing is still contested, and minimal research has been done in languages other than English. In recent years, researchers in this field have considered questions like these:

- *Is there a characteristic writing style for a certain gender?*
- *Can texts be characterized on a scale from gendered to genderless?*
- *Does the "best" writing capture a genderless, universal perspective?*

Most of the research relating to computationally characterizing gender has been done on English texts. I am interested in applying these techniques to Russian texts. Today, Russian culture is steeped in sexism and misogyny. Gender is a fraught topic in Russia, as feminism is seen as a threat, and trans and nonbinary people face serious stigma, discrimination, and danger. There is still an out-sized, politicized focus on the "traditional" family structure and government pressure for women to conform to "traditional" gender roles. In a video conference in 2019, President Vladimir Putin said:

> "In this day and age, [women] have attained the heights of practically all professions ... and at the same time you remain beautiful, charismatic, charming, the center of gravity for the whole family, uniting it with your love, with your capacity to inspire and support, to give warmth and comfort."[1]

Female voices have a history of being pushed to the margins of influence, and the rich Russian literary tradition is overwhelming comprised of male voices (think Tolstoy, Gogol, Dostoevsky). There are very few widely known Russian female novelists, and most of the famous Russian female authors are poets, which no doubt was seen as a more acceptable venture for women. But I highly doubt that Russian women abstained from writing novels and short stories. Especially in the 19th century, Russian female authors probably wrote under pseudonyms in order to publish, and it's possible that some of their texts went lost or unpublished in a male-dominated literary sphere

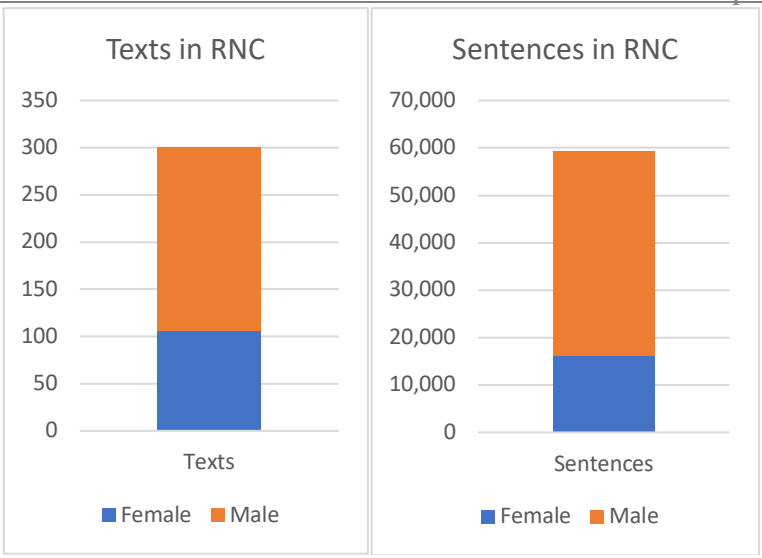
The purpose of my project is to computationally explore the writing styles of female and male Russian authors and answer some of the questions proposed above. I would also like to pay homage to Russian female writers and possibly uncover a model that would allow us to accurately classify the texts of female authors that previously went unidentified and unrecognized.

---

[1] https://apnews.com/70499d77d5bd4ea3b4462d32907420d4#:~:text=Women%20in%20Russia%20may%20hold,and%20sexual%20harassment%20have%20hardly

The dataset that I am using is the Russian National Corpus (RNC), a collection of documents that come from a variety of sources (blog posts, speeches, fiction, science articles, and journalism). The RNC is an annotated corpus, so each word is supplemented with various grammatical and linguistic features, and each document is labeled by metadata such as time of publication, genre, and gender of the author. There are 556 total texts in the corpus, and 300 of them are annotated with the author's gender. These 300 relevant articles (194 male and 106 female authored) contain 59,265 sentences (43,273 male and 15,992 female authored), which I used to run my analyses.

Male vs. female authored text in the Russian National Corpus



I separated the 59,265 sentences into 9,755 excerpts of 5 sentences (7,132 male and 2,623 female). I then applied Scikit-Learn's CountVectorizer, a bag-of-words model, and TF-IDF vectorizer to the corpus to get unweighted and weighted word counts.[2] From the CountVectorizer vectors I created LDA topic model vectors, and from the TF-IDF vectors I created NMF vectors. I looked at the top CountVectorizer and TF-IDF words across the categories.

Top words across categories:

```
count_vecs female ['как', 'это', 'я', 'а', 'с', 'что', 'на', 'не', 'в', 'и']
count_vecs male ['он', 'как', 'а', 'я', 'с', 'что', 'на', 'не', 'в', 'и']
tfidf_vecs female ['у', 'это', 'я', 'что', 'а', 'с', 'на', 'не', 'и', 'в']
tfidf_vecs male ['как', 'он', 'а', 'что', 'с', 'я', 'на', 'не', 'в', 'и']
```

These results were fairly interesting, since I noticed that the word "он" (meaning "he" in Russian) appears near the top of the list for the male authored datasets while not appearing at all in the female authored datasets. This is an interesting statistic, given that many of the male authored texts are written in the third person, which implies that there isn't an explicitly male character in the writing.

[2] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html