Unpacking Russian Presidential Speech Patterns with Machine Learning

published by Clayton Thomas Besaw on May 22, 2020

Alexander Litovchenko / Public domain

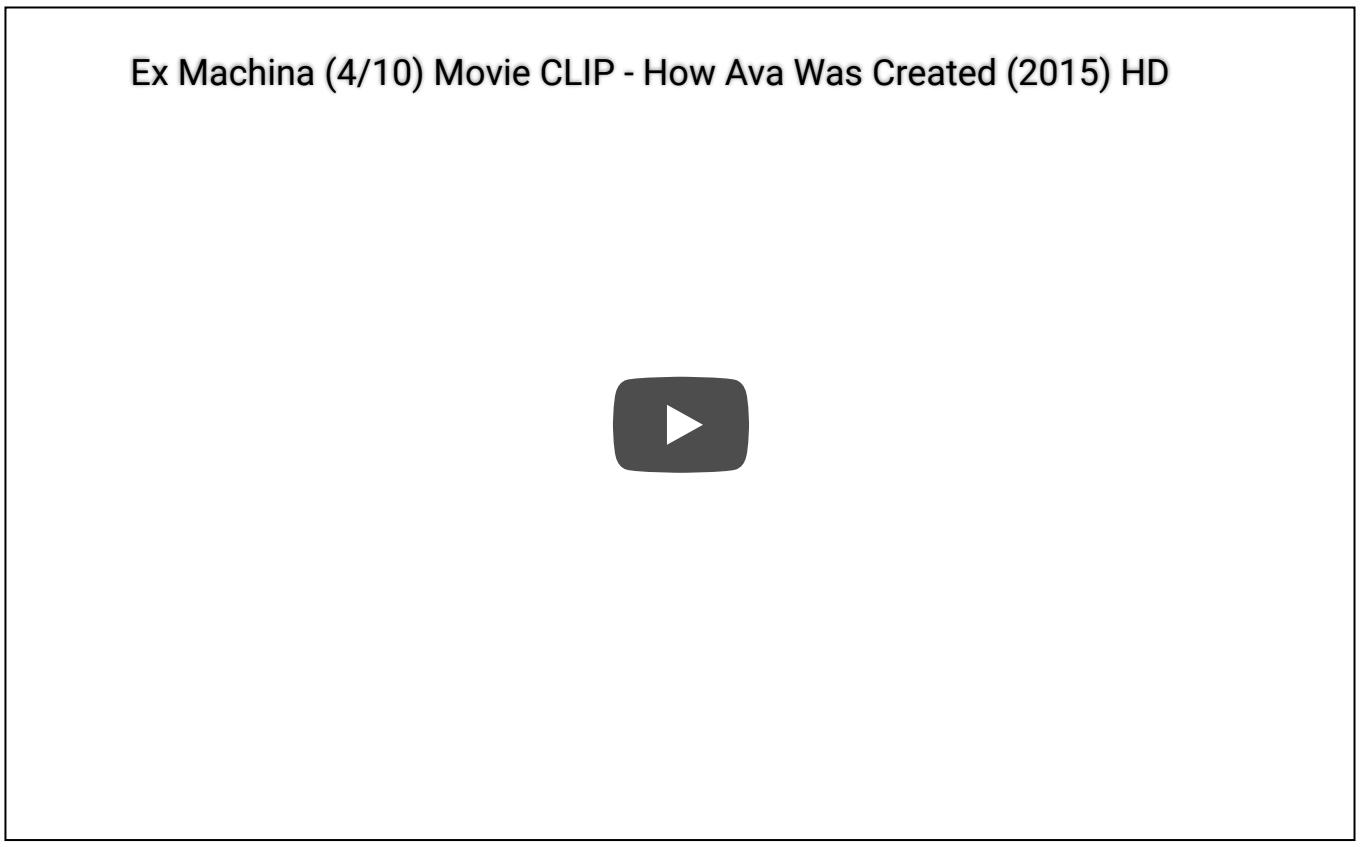
Natural language processing and topic modeling specifically have proliferated in accessibility over the past decade. Topic modeling algorithms such as Latent Dirichlet Allocation (LDA) are now considered a type of “work-horse” tool for uncovering patterns (topics) in large sets of text data.

LDA, like other topic modeling approaches, is an unsupervised machine learning technique that identifies “clusters of words” that can be grouped into discrete concepts or topics. The model does not identify these topics on its own beyond a vector of individual word association. As a result, it is up to the analyst to explore these latent topic clusters and to assess what concept(s) best describe the statistical associations uncovered by the algorithm.

So what can algorithms like LDA contribute to the world of OSINT? Simply, it can provide insight into patterns found in consequential political language.

Politics and language have been so intertwined throughout human history that it is nearly impossible to separate the two.

Language underpins our very humanity as the subjective manifestation of our cognitive styles and underlying belief systems. Essentially, language is the direct manifestation of our intelligence as human beings.



In the great sci-fi thriller Ex Machina Oscar Issac's character Nathan (a morally ambiguous tech CEO) highlights this very concept.

> *They thought that search engines were a map of what people were thinking, but actually they were a map of how people were thinking. Impulse, response, fluid, imperfect, patterned, chaotic.*
>
> *— Nathan (Ex Machina)*

It is not a secret that language is not just about the words used, but that the words used (and how they are used) can provide a window into the very psychological traits that underpin human intelligence and decision making.

Nathan Leites and his Operational Code Theory (1951) sought to provide tools for assessing the cognitive style of political leaders, specifically the Soviet Politburo as a strategy for anticipating future behavior.

Leites empirical approach has been built upon within political psychology with work on supervised classification of political speech text through Operational Code and Leadership Trait Analysis.

Non-positivist critical scholars such as Foucault and Derrida also highlight the importance of language in not only the explanation of the subject's understanding of reality, but also how language shapes and reinforces norms, ritual and power structures within society (i.e. what is taken for granted and what is not)

Given that language is strongly associated with a number of salient political and social processes within human behavior and broader societal development, it suggests that the ability to rapidly monitor and explore political texts could be highly useful for the real-time analysis of relevant political speech from an open source intelligence perspective.

While the above research fields have contributed much to the broader study of political language, they are also highly insular and access to tools and conceptual understanding is often highly contained within the academy.

Machine learning algorithms such as the LDA model cautiously allows any analyst to examine and unpack important trends within political text.

However, no single theory or model can realistically capture all of the dynamics that play out within human political speech and the analyst must rely both on their substantive expertise of the text being modeled and have confidence in their ability to run, maintain and address flaws associated with their model.

Russian presidential speeches as subject

There is no doubt that the Russian Federation under the leadership of President Putin has dramatically transformed its geo-political positioning over the last decade.

Regardless of one's belief in the current allegations surrounding Russia's involvement in domestic political discord in the United States and other western democracies, it is no secret that Putin has become increasingly hostile to the western democratic bloc. Making him an important subject for western foreign policy decision making and analysis.

He is also a leader that excels at the strategic use of political language as can be highlighted by his Holocaust forum speech given earlier this year.

Given his status and aptitude for political language, Russian presidential speech text present an ideal subject for demonstrating the strengths of the LDA model for uncovering and monitoring text topics over a leader's tenure.

So far, so good right?

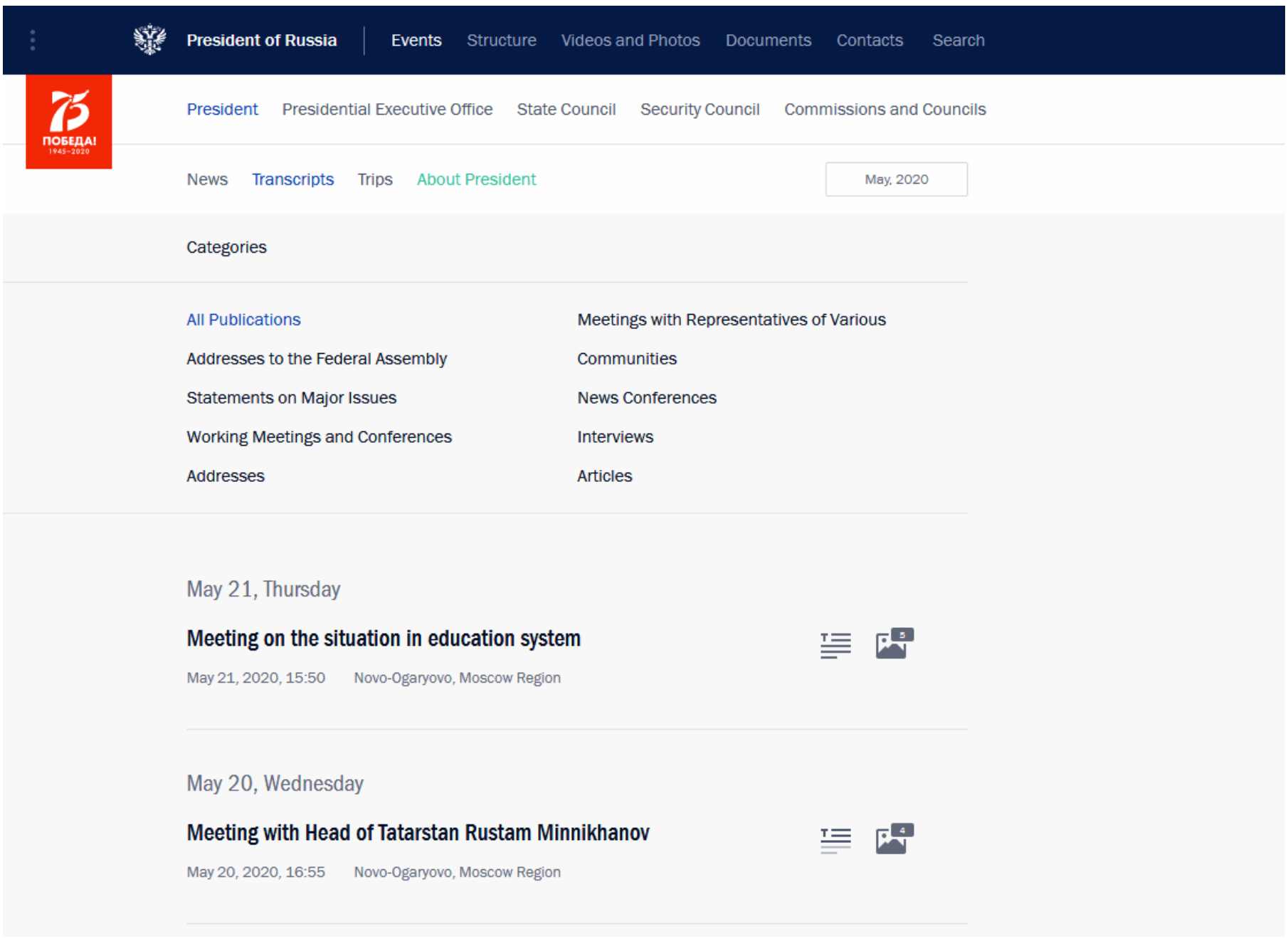
We know the subject is important, but what exactly does topic modeling help in regards to monitoring consequential political text/speech?

By using topic modeling we can specifically attempt to answer the following questions regarding official Russian presidential speech materials.

1. What kinds of latent conceptual topics are appearing in presidential speeches?
2. What topics are most prevalent?
3. What words make up each topic cluster?
4. How related are the topics?
5. How does topic presence change over time?

Unpacking latent topic patterns in Russian presidential speeches

To train our model for answering the above questions, we first collected every officail Russian presidential speech transcript up to April 23rd, 2020. This was done by building a bot that would navigate through the Kremlin's English language *Transcripts* webpage. Our scraping bot extracted information regarding the title, location, text, key-words and date of speech for 7,299 speeches in total.



Source: http://en.kremlin.ru/events/president/transcripts

Once we had our text materials scraped and placed in a data-frame, we then performed a number of text cleaning/parsing operations to make our data ready to train our LDA topic model.

These operations removed punctuation and unnecessary symbols. It also got rid of common English stop-words (e.g. the, is, at, which, on). Stop-words are often commonly used function words that have little substantive meaning in their own right.

Next we broke all of the remaining words down to their lemmas. Lemmas are the base form of a word before inflection takes place. By converting all words to their base lemma, we can make sure that we are not counting the same word multiple times due to inflection differences.

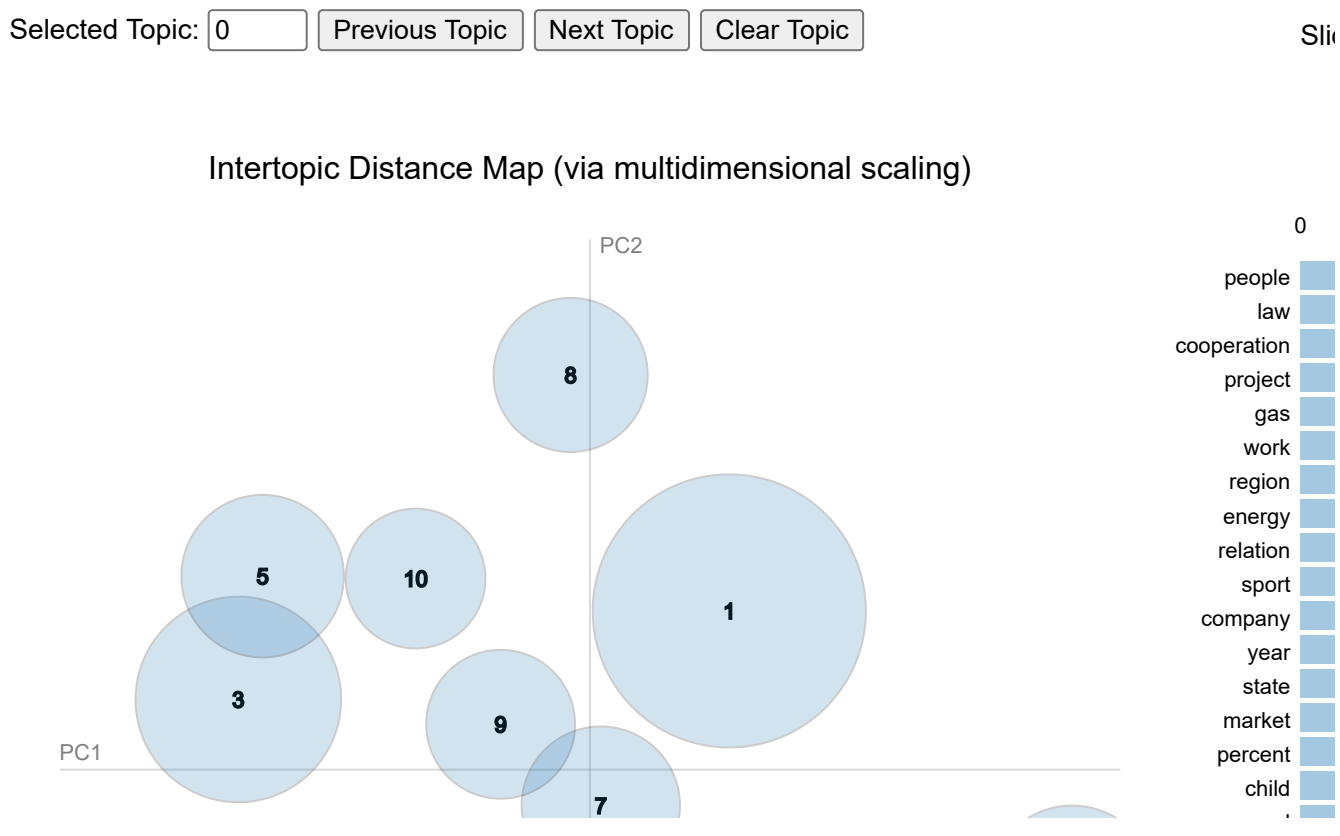
With our parsed text data, we then created a document term matrix (corpus) that would allow us to train our LDA model using the Gensim natural language processing library in Python.

For this analysis, we found that a basic ten-topic model performed reasonably well. Simply, we told our LDA model to uncover ten latent topic clusters within the speech corpus. This is done by examining the ways in which words are used in relation to each other (clustering). Our algorithm does not understand human language, but it can show us statistical patterns in how individual words are used in clusters.

The underlying idea here is that humans use words in both conscience and subconscience ways. Theoretically, our model should be able to uncover topic patterns within the text that the speech giver/writer may themselves not even fully realize. It also allows us to quickly classify such topics in both the aggregate corpus and per-document in a much more efficient way than human coding, something that takes significant time and inter-coder comparison to achieve (on top of subjective biases that may color a human analysts insights).

So what did we find once we analyzed our Russian presidential speech corpus?

Aggregate Model Insights - It's all about money and politics



10 Topic LDA Model with Putin Speech Corpus. Source: https://oefdatascience.github.io/KremLDA/index.html

Our trained LDA model has been visualized in the interactive dashboard above. On the left, you will see ten bubbles that range in both spatial position and size.

Here, size refers to the overall proportion of the topic within the aggregate corpus. The bigger the bubble, the more prevalant that topic is within our corpus of 7,299 speech texts.

Additionally we can see that some bubbles overlap and some are far apart from each other.

Bubbles that overlap share similar topic words and have some conceptual relationship with each other. Bubbles that are further apart share less underlying words like "people", "country", "problem",

Navigating to the right-hand side of the dashboard, you can see a horizontal bar chart that displays word frequency data. To explore the specific word clusters that make up each topic, simply click on a bubble and watch the words on the right change.

The red bars indicate topic specific word frequency with the blue bars representing the overall frequency within the aggregate corpus.

Exploring the words on the right also allows an analyst to make judgements about the inter-subjective meaning behind the topics uncovered by our model. We recommend that you change the relevance metric (lambda parameter) at the top right corner from 1 to .4. This allows a less conservative inclusion fo words and allows teh analyst to examine potentially more meaningful words for that specific topic.

For topic 1, we see words like "people", "party", "election" and "political". This suggests that topic 1 represents some concept clustering around domestic politics and political leadership. [1]

After examining all ten topic clusters, we identified the following topics in order of corpus presence:

1. Domestic politics
2. Diplomacy and trade
3. Economy
4. Nationalism
5. Military/Defense
6. Oil and energy
7. Regional politics
8. Crime and punishment
9. Sport and culture
10. Human development

Substantively these topic categories make a lot of sense! It makes sense that a Russian president would focus significant attention on domestic politics, diplomacy/trade and economic considerations.

Nationalist language and military/defense make up the bottom half of the top-5 topics. Oil/energy, regional politics, crime/punishment, sport/culture and human development make up the latter five topics in terms of corpus presence.

Economy and Military/Defense also share the largest overlap between any of the topics. Regional politics and sport/culture also share a smaller relative conceptual overlap.

Most of the topics also cluster around each other with the exception of topic 2 (diplomacy/trade), topic 4 (nationalism) and topic 6 (oil/energy). These three topics are more distant from the other seven topics and from each other as well. This suggests that these topics are often used in isolation and are more unique in the types of words used to construct their latent conceptual form within the text data.

From this model we can ascertain that domestic politics, diplomacy/trade and the economy are the most prominent topics uncovered in the aggregate corpus.

We can also suggest that nationalism, oil/energy and diplomacy/trade are highly unique topics within the corpus and may warrant additional consideration in the future.

Topic changes across time

Examining the aggregate findings proved to be highly interesting. We gained a sense of the kinds of latent conceptual topics being used in Russian presidential speeches and the specific words that make up the topics.

However, politics is just as much tied to the temporal domain as it is the linguistic one.

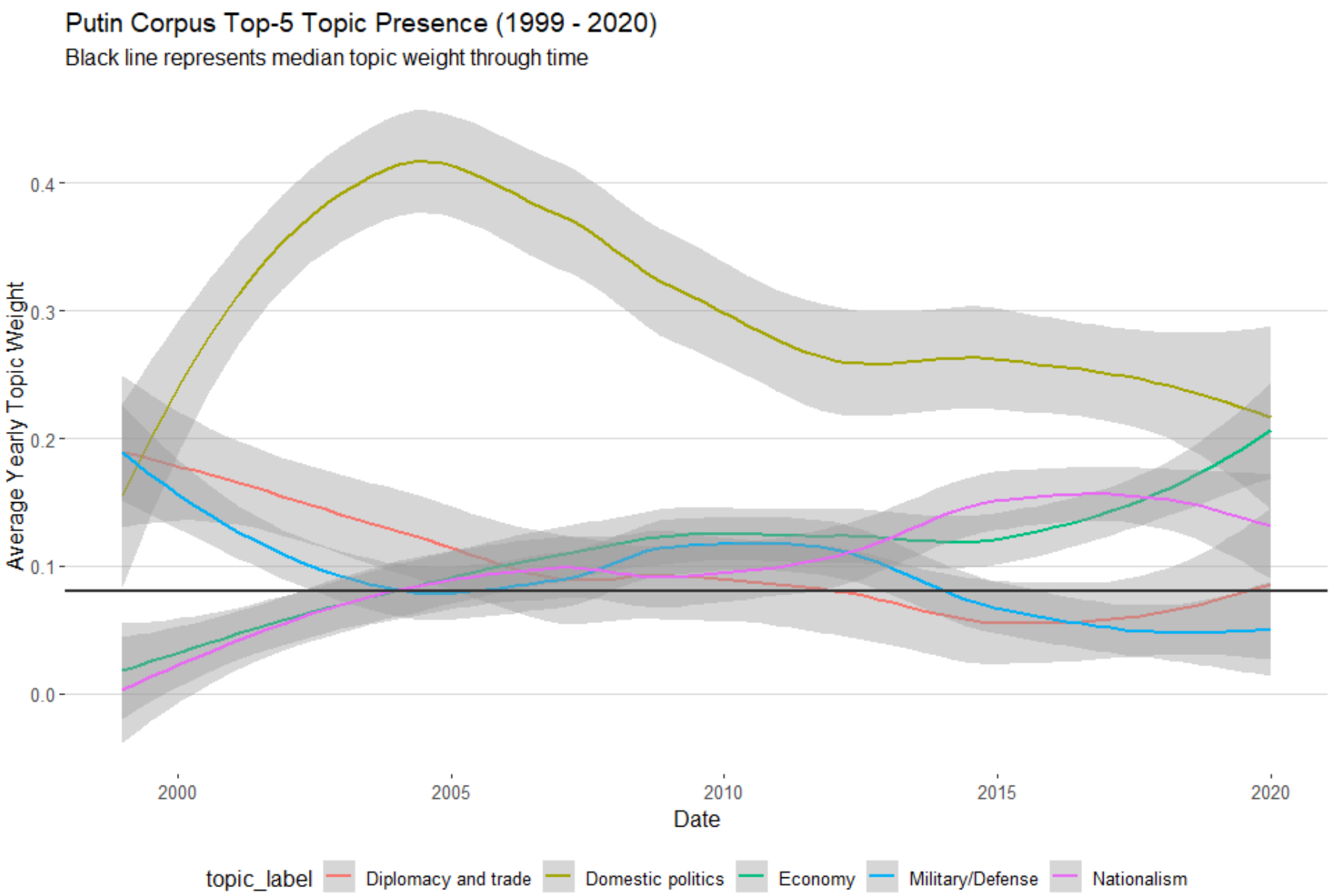
Politicians learn and change over time. They may use different language or change priorities based on experience or constraints in that time period.

Unfortunately our dashboard cannot give us insight into how topic presence changed over the history of our corpus, but this wont stop us from finding ways to calculate such patterns!

Because our model allows us not only to examine aggregate trends, but to also quantify topic manifestation by individual documents, we can calculate yearly weighted averages regarding topic proportions for every year in the corpus (1999 - 2020).

After extracting document-level topic proportion data, we simply calculated a "year proportion weight" for every topic based on document-level information for all documents within a specific year.

The result is a data-set that allows us to visualize topic trends across the entire temporal period of our corpus.



Data Source: https://github.com/OEFDataScience/KremLDA/blob/master/year_topic_weights2.csv

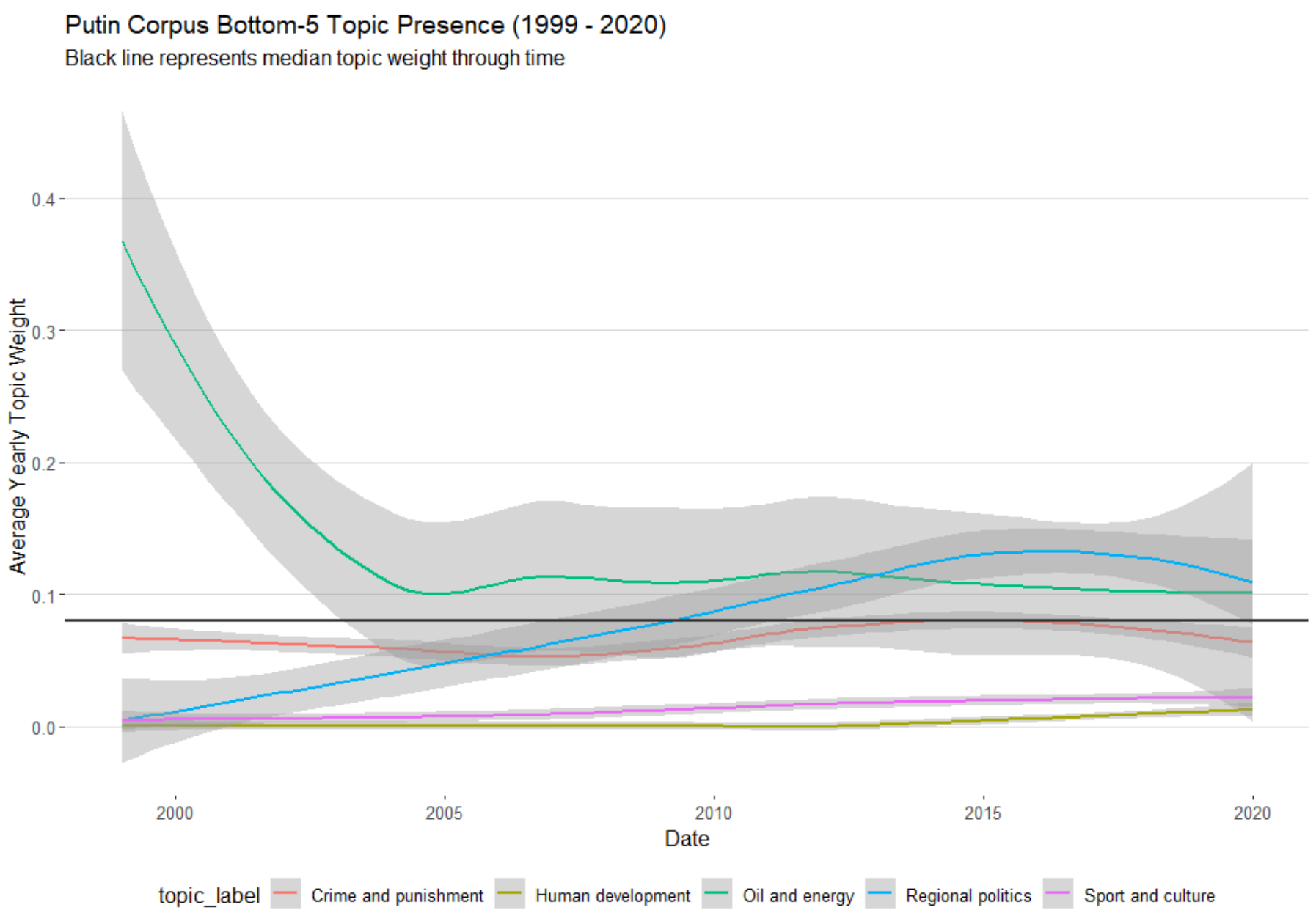
Here we can see a number of interesting patterns regarding the top-5 topics already. The solid black line represents the median topic weight within the corpus and can be a useful marker for demonstrating when a topic is above or below the central tendency for representation.

As expected, domestic politics dominates for much of the the last two decades. It peaks in 2004 before steadily declining over the next 15 years. As of 2019-2020 though, domestic politics begins to see parity with the economy and nationalism.

Diplomacy/trade and military/defense started out above the median weight, but has slowly declined overtime with both being at or below the central tendency as of 2019-2020.

In contrast, both nationalism and the economy started out well below the median weight and both have steadily climbed in presence over the past two decades.

Overall, we can see that domestic politics has been the predominant topic over time while nationalism and the economy have become steadily more present in recent years.



Data Source: https://github.com/OEFDataScience/KremLDA/blob/master/year_topic_weights2.csv

Examining topics 6-10, we find interesting but unsurprising patterns.

OIl/energy is by far the most represented of these topics within the corpus overtime. It is always above the central tendency, but has plateaued slightly above the median topic weight since around 2004. As such it is a constant but only slightly above average topic in terms of presence over time.

Regional politics in contrast has risen like a phoenix from obscurity over the past two decades. This topic became more present starting in 2009-2010 and has now overtaken oil/energy in terms of overall presence within Russian presidential speech patterns.

Crime/punishment is the only other topic to reach central tendency, but has been consistently below average across the entire temporal period captured here.

Finally, sport/culture and human development are unsurprisingly at the bottom. Both were last and second to last in terms of topic presence in our aggregate dashboard and that trend stays the same over the past two decades.

Final Takeaways

So what can we conclude from this analysis?

Russian presidential speech material is predominately focused on domestic political considerations, trade and the economy.

Additionally, domestic political considerations have been the predominant structural concept underlying most speeches across the last two decades until the last two years. Since 2019 though, nationalism and the economy have risen to near parity with domestic politics as manifested topics within Russian presidential speeches.

This information allows us to not only explore past patterns in speech by Russian presidents, but it also allows us to update and monitor new speeches in real-time.

With our trained model and additional analytics, we can further train our LDA algorithm with each new speech or ask our model to classify new speeches accordingly.

This allows us to almost instantaneously monitor and classify new speech materials for their topic proportions and to further

update our time-series visuals to examine potential shifts in topic patterns as they happen in real-time.

While no specific methodology can ever be perfect or all-encompassing, this analysis demonstrates that topic modeling can provide an accessible and useful framework for producing open source intelligence regarding not only what important political leaders are saying but also provides insight into how latent speech topics (1) relate to each other, (2) how "present" each topic is and (3) how these topics change in "presence" over the lifespan of a regime or an individual leader's tenure in office.

With this model and data we can thus explore not only how this use of political language has manifested overtime, but also create a real-time monitoring pipeline that allows us to classify novel political language as it is released to the analyst.

Thank you for taking the time to read this analysis and hopefully it can inspire further work on how machine learning models can contribute to the rich tapestry of OSINT production and the analysis of foreign policy decision making.

[1] If you are a curious analyst yourself, click through all of the topics before reading further. Write down your interpretation of each topic and compare them to ours!

Machine Learning Data Science Russia Natural Language Processing Geo-politics Political Speech Text Analysis Putin OSINT

Share this on :