Repository: https://github.com/AlvaroGM88/Matching-function-

Abstract

This report aims to explain the steps followed to create a function that matches Bloc Center faculty research with different text (bills, think tanks articles, among others) where the main topic is public policy issues related with artificial intelligence.

For the text of interest, the plain text is used, while for the research of the Bloc Center faculty, keywords from different research papers were used.

The pre-process follows the standard steps used in this kind of task, like filter numbers, lemmatized text, and lower case all the words in the text. Then tf-idf function is used to vectorize the inputs (text and keywords), and the Euclidean distance to calculate the similarity between research and text of interest. the final output of the function is an excel that for a given text of interest give the five closest scholar base on the Euclidean distance. Finally, the report identifies several recommendations to improve this first version of the matching function.

The motivation of this summer project is to achieve a correct match between scholars participating in the Bloc Center and bills related to artificial intelligence currently under discussion in the US Congress, in order to eventually be able to influence the future law derived from this bill, through expert judgment. This does not make it exclusive that the matching between faculty research and public policy discussions should be only bills, but other text can be included, such as articles written by think tanks.

It should be noted that, given the nature of the problem, this falls under the category of unsupervised learning, which makes it challenging to evaluate the tool, the before mentioned makes it necessary to consult with Bloc Center faculty if they consider the matching to be correct, a process that was not carried out during this first version.

To achieve the above is necessary to define three main variables:

i. The bill or text that would be of interest to the Bloc Center participant.
ii. A summary of the main topics investigated by the scholar.
iii. A function that can match the variables defined previously in the points i. and ii.

Bill or text of interest

At this stage of the project, three types of text were used that could be of interest: 116th congress bills that were related to artificial intelligence, the same for 115th congress bills and all up-to-date articles published in Brookings under its Blueprint for the future of AI initiative[1]. The files were transformed to .txt format and reviewed one by one in order to correct minimal problems in the text resulting from the conversion to the new format. Modified texts can be downloaded on the web page of the repository. The detail of the texts with a brief description can be found in annex of this document

Summary of topics investigated by Bloc Center faculty

For the construction of these variables, two methods were chosen. The first was the construction of a dictionary of keywords for a given professional, according to the research published in the different pages of the facultative, taking on average between 4 to 6 works for each researcher (this was a manual work). The second approach was using all the keywords of the research publications collected by CMU elements. It is important to note that not all the research of the scholars are in this document, since they either do not have keywords or have not yet been identified by CMU elements, in addition there is an obvious imbalance in the information available to each academic and the absence of some professors identified previously, either because they do not belong to CMU or are not in the CMU elements records.

Matching function

The matching function was developed in Python (reference Match_func.py). The file decomposes into 4 different chunks. The first call the libraries and specific functions that will be used later. The second corresponds to defining the path of the folder with the texts of interest (.txt format), identifies the excel file with the keywords of each scholar and, if necessary, set the working directory where the output will be saved. The third is the matching function that will be detailed in the next section and, the last chunk only runs the matching function, given the inputs defined in chunk two.

[1] https://www.brookings.edu/series/a-blueprint-for-the-future-of-ai/page/2/

Explaining def matching[2]

The matching function was defined mainly as a function with inside functions, where each inside function accomplishes a specific task to process the data before the pairing. The description of each function is as it follows.

get_bills_text: extract the .txt from the file already define in data_folder and assigns a label that is the file title, for each different document. The output of the function is a list of the text of the material in the data_folder and a list of labels (names) for each of those documents.

Elem_keywords: extract the information of faculty and keywords provided by CMU elements from an excel file. The function groups by the name of faculty all the keywords defined in the column 'Keywords.' It is important to notice for this function to work correctly it is necessary to identify in the queried excel file a column named 'Faculty Name' with the professor name (or other identification) and another column 'Keywords' with the keywords for each research paper. The output of the function is a dictionary that assigns an identifier (key) number to each of the academics and list of keywords for each one.

is_letter_only: applied the function "isalpha," that filter the text and only leave words (exclude numbers) in the different documents and keywords.
For the next steps, a set call all_names is defined to be used to filter the names from the texts and the lemmatizer function to preprocess the texts.

clean_text: clean all the inputs applying the preprocess function defined before and applied function "lower," to set all words in the text to lowercase, returns a list of the text "clean."

For vectorizing the different text to compare the similarity between documents and keywords, frequency-inverse document frequency (tf-idf[3]) function is used to fit in the text of interest (example: bills) and then the keywords. The output is a set of matrices that for each scholar and text of interest give a score (define by tf-idf) for the words used in the inputs.

Then a loop computes the Euclidean distance between keywords of each scholar and all the different text of interest; the next step ranks for each text all the researchers base on the distance previously calculated (ascending), the final output of the overall function is an excel file that for a given text retrieve the first 5 ranked professors.

In conclusion, this first approach to the problem of how to match the text of interest and faculty from Bloc Center delivers a function that effectively achieves that purpose, but it is only a first version. It is necessary to ask the scholars if the results make sense to them and try other approaches like the one suggested in the recommendation section, to effectively achieve the best method possible to this task.

[2] Yuxi (Hayden) Liu. Python Machine Learning by Example. Second Edition. Chapter 2 and 3.

[3] The details of this specific function can be found in the page 97 of Yuxi Liu book.

Recommendations

- For texts that are of interest to the Bloc Center faculty, take a summary made by experts and use it as input. This would allow better filtering of the wording in the texts of interest, which could improve the matching.
- For the keywords for each scholar, ask them to record which are the topics that should be considered in these keywords. This will also make it possible to incorporate scholars that are not included in the records of CMU elements.
- Try other stemming functions instead of lemmatizing function.
- Tweak parameters used in the tf-idf function and try bigrams (currently the function is set in unigram), for this is necessary to discuss with professors if the match was correct.
- Try another measure of distance like Manhattan or Chebyshev.

Annex

116th Congres Bill

Bill	Title	Description
hr1367	Children and Media Research Advancement Act’’	To amend the Public Health Service Act to authorize a program on children and the media within the National Institutes of Health to study the health and developmental effects of technology on infants, children, and adolescents.
hr1485	Computer Science for All Act of 2019	To authorize the Secretary of Education to carry out a program to increase access to prekindergarten through grade 12 computer science education.
hr153	Supporting the development of guidelines for ethical development of artificial intelligence	Supporting the development of guidelines for ethical development of artificial intelligence
hr1668	Internet of Things Cybersecurity Improvement Act of 2019	To leverage Federal Government procurement power to encourage increased cybersecurity for Internet of Things devices, and for other purposes.
hr2202	Growing Artificial Intelligence Through Research Act’’	To establish a coordinated Federal initiative to accelerate artificial intelligence research and development for the economic and national security of the United States, and for other purposes.
hr2231	Algorithmic Accountability Act of 2019’	To direct the Federal Trade Commission to require entities that use, store, or share personal information to conduct automated decision system impact assessments and data protection impact assessments.
hr2432	Future Defense Artificial Intelligence Technology Assessment Act	To require the Secretary of Defense to submit a report to Congress on the artificial intelligence strategy of the Department of Defense.
hr2542	Preparing Localities for an Autonomous and Connected Environment Act	To direct the Secretary of Transportation to make grants for the operation of a clearinghouse to collect, conduct, and fund research on the influences of highly automated vehicles on land use, urban design, transportation, real estate, and municipal budgets, and for other purposes.
hr827	Artificial Intelligence Job Opportunities and Background Summary Act of 2019	To promote a 21st century artificial intelligence workforce.
s1363	AI in Government Act of 2019	To authorize an AI Center of Excellence within the General Services Administration, and for other purposes.

s1398	Smart Cities and Communities Act of 2019	To promote the use of smart technologies and systems in communities, and for other purposes.
s1471	Armed Forces Digital Advantage Act	To require digital engineering as a core competency of the Armed Forces, and for other purposes.
s734	Internet of Things Cy5 bersecurity Improvement Act of 2019	To leverage Federal Government procurement power to encourage increased cybersecurity for Internet of Things devices, and for other purposes.
s847	Commercial Facial Recognition Privacy Act of 2019	To prohibit certain entities from using facial recognition technology to identify or track an end user without obtaining the affirmative consent of the end user, and for other purposes.

115th Congres Bill

Bills	**Title**	**Description**
hr4625	Fundamentally Understanding The Usability and Realistic Evolution of Artificial Intelligence Act of 2017	To require the Secretary of Commerce to establish the Federal Advisory Committee on the Development and Implementation of Artificial Intelligence, and for other purposes.
hr4829	Artificial Intelligence Job Opportunities and Background Summary Act of 2018	To promote a 21st century artificial intelligence workforce.
hr5356	National Security Commission Artificial Intelligence Act of 2018	To establish the National Security Commission on Artificial Intelligence.
s3502	AI in Government Act of 2018	To authorize an emerging technology policy lab within the General Services Administration, and for other purposes

Brookings AI

Identifier	Title	Description
adjust_automation	How to adjust to automation	President Clinton’s first inaugural address contained a phrase that is as relevant today as it was then: “the urgent question of our time is whether we can make change our friend and not our enemy.” It’s a phrase that he never used again, as far as I can tell, any other time over his eight years in office.
AI_America_dig_city	Artificial intelligence in America’s digital city	Cities are an engine for human prosperity. By putting people and businesses in close proximity, cities serve as the vital hubs to exchange goods, services, and even ideas. Each year, more and more people move to cities and their surrounding metropolitan areas to take advantage of the opportunities available in these denser spaces.
AI_education_workforce	The role of AI in education and the changing US workforce	The growth of artificial intelligence (AI) and emerging technologies (ET) is poised to reshape the workforce.1 While the exact impact of AI and ET is unclear, experts expect that many jobs currently performed by humans will be performed by robots in the near future, and at the same time, new jobs will be created as technology advances. These impending changes have important implications for the field of education. Schools must prepare students to remain competitive in the labor market, and postsecondary institutions must provide students and displaced workers with relevant education and retraining opportunities. Innovations in technology will also create new tools to support educators, students, and others seeking retraining and employment.
AI_ethical_dilemmas	The role of corporations in addressing AI’s ethical dilemmas	The world is seeing extraordinary advances in artificial intelligence. There are new applications in finance, defense, health care, criminal justice, and education, among other areas.1 Algorithms are improving spell-checkers, voice recognition systems, ad targeting, and fraud detection.
AI_financial_consumers	How artificial intelligence affects financial consumers	Artificial intelligence (AI) technology has transformed the consumer financial services market and how consumers interact with the financial services ecosystem. This paradigm shift has been driven by the accelerated maturation of the algorithms; the historic level of investment flooding the financial services market; the competition for market share between incumbents and new entrants; and rapid changes in consumers’ preferences for digital financial products. From AI-driven chatbots to sophisticated wealth robo advisors, AI applications have clear potential to expand opportunities for consumers living at the margin. However, experts have yet to discuss the relevance of AI for consumer financial protection in earnest, including the implications of AI solutions that could better protect consumers.
AI_future_warfare	The role of AI in future warfare	To illustrate how artificial intelligence (AI) could affect the future battlefield, consider the following scenario based on a future book I am writing entitled The Senkaku Paradox: Risking Great Power War over Limited Stakes. The scenario, imagined

		to occur sometime between now and 2040, begins with a hypothesized Russian “green men” attack against a small farming village in eastern Estonia or Latvia. Russia’s presumed motive would be to sow discord and dissent within NATO, weakening the alliance. Estonia and Latvia are NATO member states, and thus the United States is sworn to defend them. But in the event of such a Russian aggression, a huge, direct NATO response may or may not be wise. Furthermore, the robotics and AI dimension of this scenario, and a number of others similar to it, will likely get more interesting as the years go by.
AI_healthcare	The opportunities and challenges of data analytics in health care	Data analytics tools have the potential to transform health care in many different ways. In the near future, routine doctor’s visits may be replaced by regularly monitoring one’s health status and remote consultations. The inpatient setting will be improved by more sophisticated quality metrics drawn from an ecosystem of interconnected digital health tools. The care patients receive may be decided in consultation with decision support software that is informed not only by expert judgments but also by algorithms that draw on information from patients around the world, some of whom will differ from the “typical” patient. Support may be customized for an individual’s personal genetic information, and doctors and nurses will be skilled interpreters of advanced ways to diagnose, track, and treat illnesses. In a number of different ways, policymakers are likely to have new tools that provide valuable insights into complicated health, treatment, and spending trends.
AI_international_trade	The impact of artificial intelligence on international trade	Artificial intelligence (AI) stands to have a transformative impact on international trade. Already, specific applications in areas such as data analytics and translation services are reducing barriers to trade. At the same time, there are challenges in the development of AI that international trade rules could address, such as improving global access to data to train AI systems. The following provides an overview of some of the key AI opportunities for trade as well as those areas where trade rules can help support AI development.
AI_National_Security_strategy	The implications of artificial intelligence for national security strategy	Artificial intelligence is transforming the world, as Brookings President John Allen and Vice President Darrell West describe in their thoughtful piece on this topic. From how we educate our youth to how economies operate, there exists no shortage of arenas where experts believe artificial intelligence will have an outsized impact. National security strategy is among them.
bigdata_AI_globaldevelopment	Using big data and artificial intelligence to accelerate global development	When U.N. member states unanimously adopted the 2030 Agenda in 2015, the narrative around global development embraced a new paradigm of sustainability and inclusion—of planetary stewardship alongside economic progress, and inclusive distribution of income. This comprehensive agenda—merging social, economic and environmental dimensions of sustainability—is not supported by current modes of data collection and data analysis, so the report of the High-Level Panel on the post-2015 development agenda called for a “data revolution” to empower people through access to information.
Credit_denial_AI	Credit denial in the age of AI	Banks have been in the business of deciding who is eligible for credit for centuries. But in the age of artificial intelligence (AI), machine learning (ML), and big data, digital technologies have the potential to transform credit allocation in positive as well as

		negative directions. Given the mix of possible societal ramifications, policymakers must consider what practices are and are not permissible and what legal and regulatory structures are necessary to protect consumers against unfair or discriminatory lending practices.
Democracy_AI	Malevolent soft power, AI, and the threat to democracy	In the space of less than a decade, the world of social media has gone from being an enabler of to a threat to democracy. While the internet can still mobilize large numbers of people to political action, it can also spew false information about candidates, suppress the vote, and affect the voter rolls and the election machinery of the state. By 2016, social media had become a weapon against democracy as opposed to a tool for democracy. Unless we are vigilant, the new world of artificial intelligence (AI) has the potential to be an even more dangerous weapon in the years ahead. This paper will look at Russian interference in the 2016 election with an emphasis on intra-party disruption and then it will look at the ways in which AI can further disrupt democracy if we are not prepared.
education_AI	Why we need to rethink education in the artificial intelligence age	Artificial intelligence (AI) and emerging technologies (ET) are poised to transform modern society in profound ways. As with electricity in the last century, AI is an enabling technology that will animate everyday products and communications, endowing everything from cars to cameras with the ability to interact with the world around them, and with each other. These developments are just the beginning, and as AI/ET matures, it will have sweeping impacts on our work, security, politics, and very lives.
energy_climate_AI	How artificial intelligence will affect the future of energy and climate	In a 2017 article for Foreign Affairs, Kassia Yanosek and I advanced the hypothesis that the biggest impacts of the information technology (IT) revolution may be felt far outside IT—in the traditional industries of oil, gas, and electricity.1 That's because IT was transforming how those industries function. That logic of transformation may be especially profound when looking at a subset of the IT revolution: artificial intelligence (AI).
gov_shape_the_future_AI	Why the government must help shape the future of AI	Rapid advances in artificial intelligence (AI) are raising serious ethical concerns. For many workers who have not seen significant wage growth in decades, AI represents a potential threat to the jobs on which they depend, and its potential interaction with the effects of globalization is alarming. Thoughtful observers worry about its capacity to intensify concentrations of public and private power, increase information asymmetries, and diminish transparency—all at the expense of citizens. In these circumstances, the significance of individual consent—one of the hallmarks of a free society—is called into question.
India_future_AI	Harnessing the future of AI in India	The size of the AI sector in India is difficult to determine, given that a lot of AI applications are in intermediary phases of production. Globally, one popular means of measuring the size of AI sectors is by adding up private sector investment in AI start-ups. According to one estimate, total AI funding worldwide has increased from $862 million in 2012 to $6.4 billion in 2017.1 The Indian AI sector, too, has seen growth in this period, with a total of $150 million invested in more than 400 companies over the past five years.2 Most of these investments have come in the last two years, when investment nearly doubled from $44 million in 2016 to $77 million in 2017

K-12_placement_algorithms	The opportunities and risks of K-12 student placement algorithms	How students are assigned to schools is changing, especially in urban areas. After decades of using students' home addresses to determine school assignments, many U.S. cities are now turning to placement algorithms—alongside school choice policies—to determine which students can attend which particular schools. These algorithms, built on the Nobel Prize-winning theory of market design, elicit families' ranked preferences for schools and use those preferences, along with schools' priorities, to match students and schools.
place_disparities_AI	Countering the geographical impacts of automation: Computers, AI, and place disparities	The 2016 presidential election revealed—as nothing before it—one of the most striking but least-anticipated aspects of the global digital revolution. In a single dramatic vote, the victory of Donald Trump highlighted the emergence of a stark and widening divide between two Americas: one based in large, digitally oriented metropolitan areas; the other found in lower-tech smaller cities, towns, and rural areas.1 In doing so, the vote displayed—with its stark red-blue map—the underrated power of technology to reshape the geography of nations.
Russia_AI_warfare	Weapons of the weak: Russia and AI-driven asymmetric warfare	Artificial intelligence is the future, not only for Russia, but for all humankind. It comes with colossal opportunities, but also threats that are difficult to predict. Whoever becomes the leader in this sphere will become the ruler of the world
US_China_AI_age	US-China relations in the age of artificial intelligence	Under President Donald Trump, great power competition has become the organizing principle of American foreign policy. This has led to near-daily invocations of the Cold War to describe the intensifying rivalry between the United States and China, and to frequent analogies to an "arms race" to describe bilateral competition in advanced technologies, including quantum computing and artificial intelligence (AI). Public statements and national plans from both governments have reinforced this zero-sum dynamic. Such framing has done more to conceal than clarify and, if taken to its logical end-point, will do more harm than good for the United States.
What_is_AI	What is artificial intelligence?	Few concepts are as poorly understood as artificial intelligence. Opinion surveys show that even top business leaders lack a detailed sense of AI and that many ordinary people confuse it with super-powered robots or hyper-intelligent devices. Hollywood helps little in this regard by fusing robots and advanced software into self-replicating automatons such as the Terminator's Skynet or the evil HAL seen in Arthur Clarke's "2001: A Space Odyssey," which goes rogue after humans plan to deactivate it. The lack of clarity around the term enables technology pessimists to warn AI will conquer humans, suppress individual freedom, and destroy personal privacy through a digital "1984."
What_is_machine_learning	What is machine learning?	In the summer of 1955, while planning a now famous workshop at Dartmouth College, John McCarthy coined the term "artificial intelligence" to describe a new field of computer science. Rather than writing programs that tell a computer how to carry out a specific task, McCarthy pledged that he and his colleagues would instead pursue algorithms that could teach themselves how to do so. The goal was to create computers that could observe the world and then make decisions based on those observations—to demonstrate, that is, an innate intelligence.