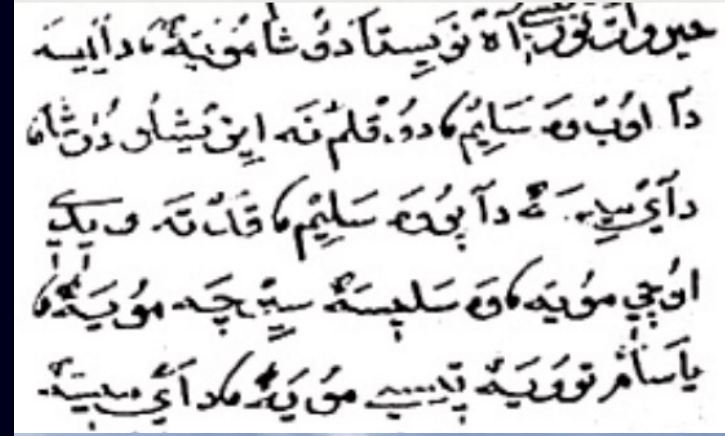
Automatic Language Identification and Relatedness Mapping

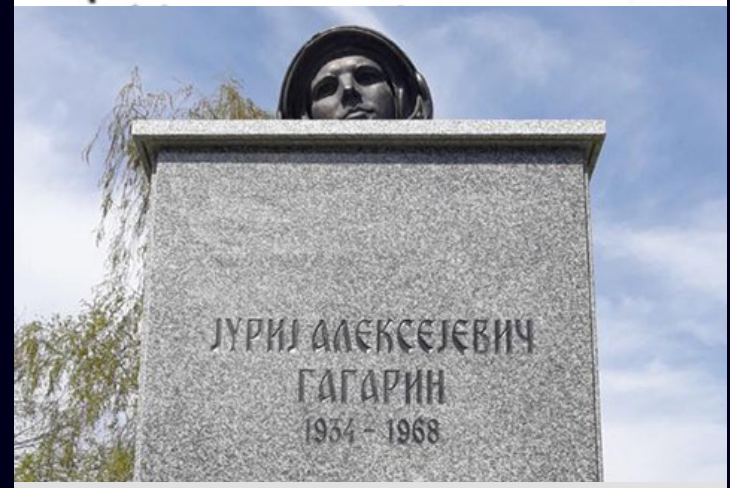
Sonia Cromp

Goals

- Language identification

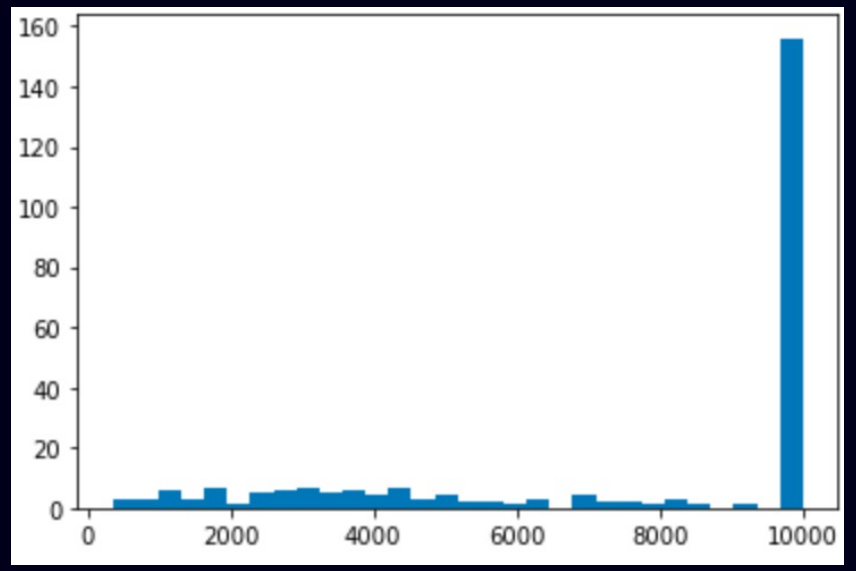
- Language relatedness
  - Ancestral
  - Geographic
  - Cultural
  - Common writing system





Dataset

- 248 languages' Wikipedia article texts
- Cleaning: Remove punctuation, non-text characters
- Result: Random 500-character "chunks"



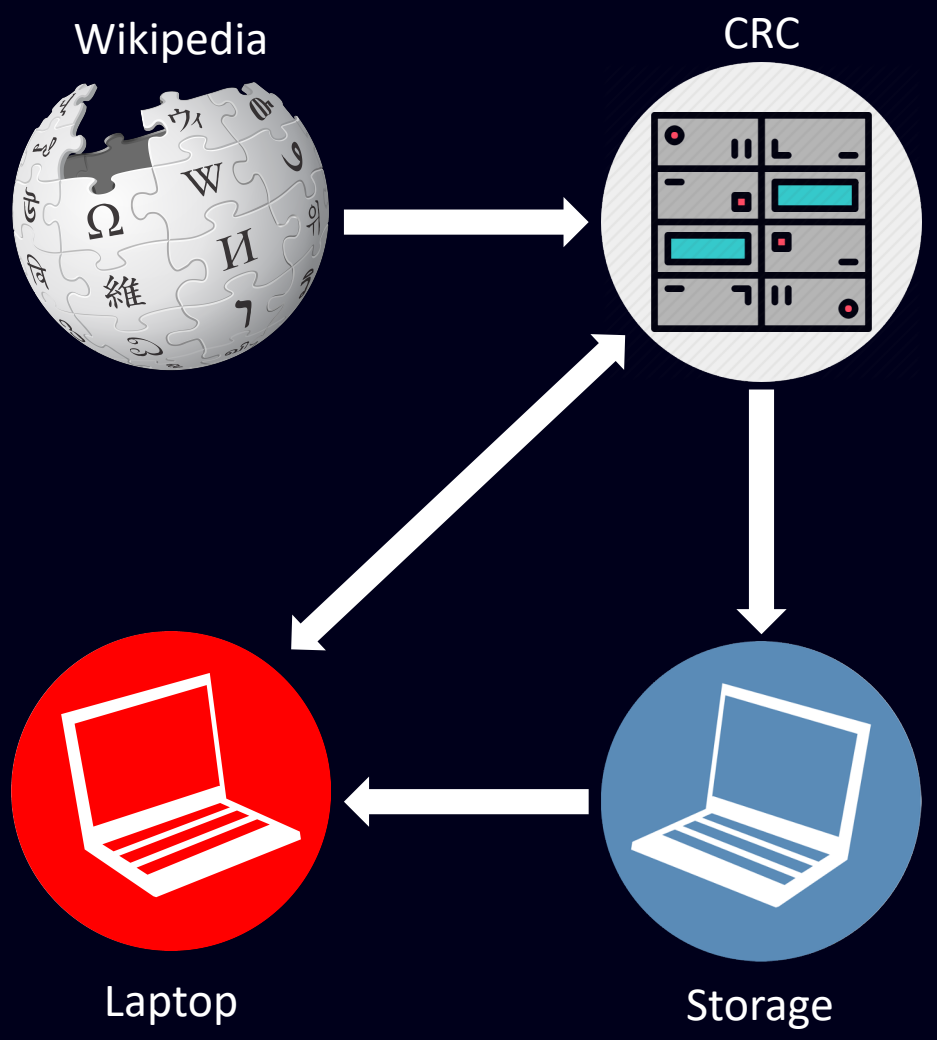
Chunks per Language

Acknowledgement: Compute resources from University of Pittsburgh

Center for Research Computing (CRC) offered through LING 1340 coursework

Data gathering

1. Networking
2. Download server dumps
3. Remove formatting, clean text
4. Chunking and shuffling



Data gathering

Before (random section of Qaraqalpaqsha)

' <timestamp>2020-05-01T08:16:04Z</timestamp>\n', ' <contributor>\n', ' <ip>185.163.26.25</ip>\n', ' </contributor>\n', ' <model>wikitext</model>\n', ' <format>text/x-wiki</format>\n', ' <text bytes="7440" xml:space="preserve">\'\'\'A\'jiniyaz Qosıbay ulı\'\'\' (a\'debiy laqabı \'\'Ziywar\'\') - XIX a\'sirdegi qaraqalpaq klassik a\'debiyatının\' en\' o\'rnekli wa\'killerinin\' biri. Ol do\'retiwshiliginde o\'zine ta\'n o\'zgeshelikke iye, og\'ada talantlı, oqımıslı, medreseni ayrıqsha tamamlag\'an aqun, ulama, ko\'rkem so\'z sheberi sıpatında basqalardan ayrılıp turatug\'ın uqıplılıq penen o\'z da\'wirinin\' progressiv …',

After (one English chunk)

'alism among other libertarian socialist economic theories As anarchism does not offer a fixed body of doctrine from a single particular worldview many anarchist types and traditions exist and varieties of anarchy diverge widely One reaction against sectarianism within the anarchist milieu was anarchism without adjectives a call for toleration and unity among anarchists first adopted by Fernando Tarrida del Mármol in in response to the bitter debates of anarchist theory at the time Belief in poli\n'

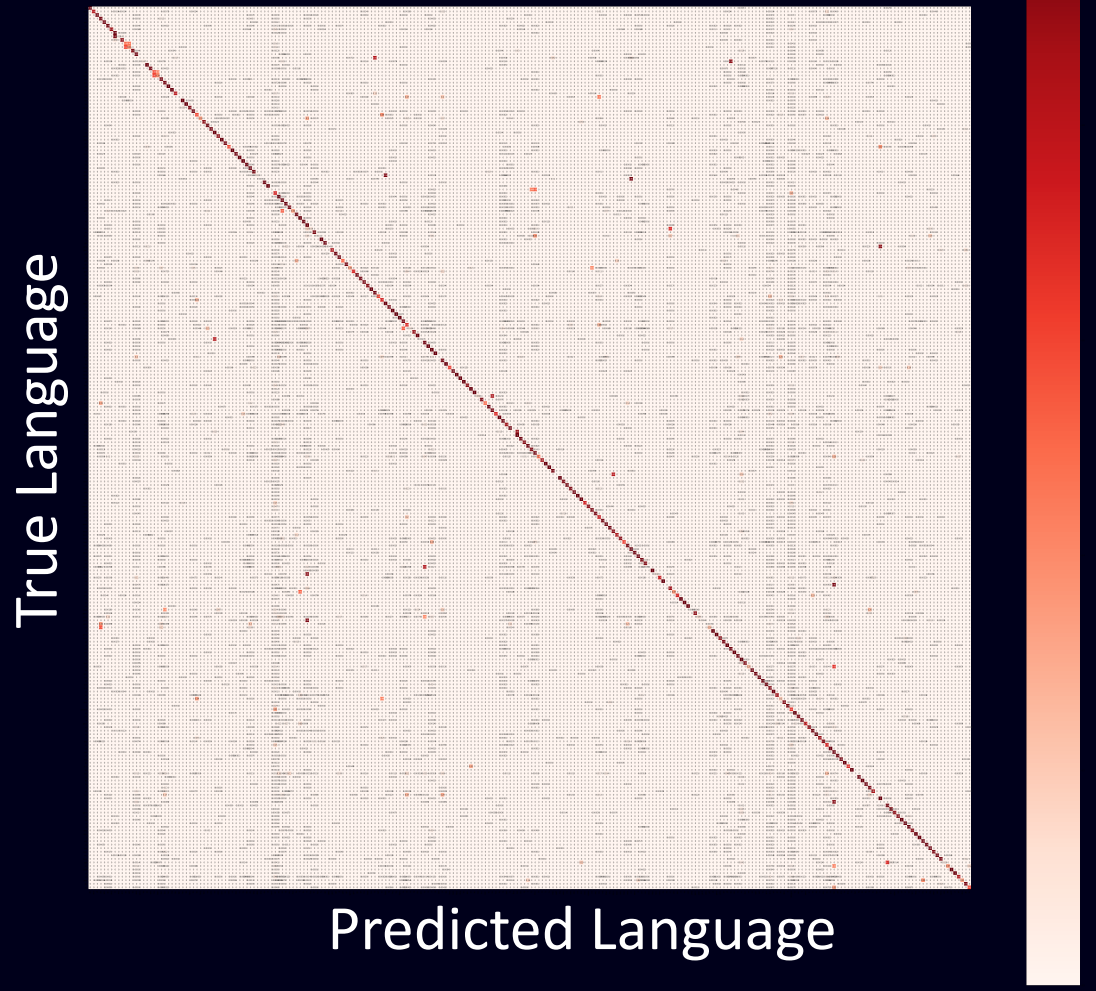
248 languages included

Kreyòl ayisyen, Limburgs, Igbo, tetun, 吴语, коми, الدارجة, latviešu, la .lojban., भोजपुरी, мокшень, Zeêuws, o‘zbekcha/ўзбекча,
تۆرکجه, සිංහල, norsk nynorsk, português, Bahasa Melayu, अवधी, 中文, тыва дыл, svenska, башҡортса, മലയാളം, Tagalog,
arpetan, slovenščina, Afrikaans, Jawa, Yorùbá, Mirandés, ગુજરાતી, Frysk, Māori, Xitsonga, sicilianu, italiano, Aymar aru, Gagauz, ????
???, davvisámegiella, føroyskt, తెలుగు, brezhoneg, català, Адыгэбзэ, dansk, 贛語, Pälzisch, Nāhuatl, Novial, qırımtatarca, Simple
English, dolnoserbski, ܐܪܡܝܐ, Oromoo, অসমীয়া, English, lumbaart, বাংলা, پنجابی, hrvatski, Nouormand, ລາວ, ଓଡ଼ିଆ, русский, 한국어,
interlingua, Papiamentu, къарачай-малкъар, Кыргызча, Kotava, български, azərbaycanca, vepsän kel’, Alemannisch, Gaelg,
Plattdüütsch, livvinkarjala, Soomaaliga, Piemontèis, Sunda, Ænglisc, galego, Hausa, Mìng-dĕ̤ng-ngṳ̄, ייִדיש, کوردی, emiliàn e rumagnòl,
ਪੰਜਾਬੀ, авар, walon, română, Picard, kurdî, مصرى, گیلکی, suomi, မြန်မာဘာသာ, kaszëbsczi, Bahasa Indonesia, ၽႃႇသႃႇတႆး , vèneto,
eesti, sardu, Kapampangan, удмурт, олык марий, 日本語, Ελληνικά, саха тыла, Runa Simi, Deitsch, bosanski, ߒߞߏ, español,
سنڌي, кырык мары, aragonés, Ilokano, Banjar, Kinyarwanda, ភាសាខ្មែរ, 客家語/Hak-kâ-ngî, српски / srpski, Wolof, Türkçe, беларуская,
Scots, Fiji Hindi, Чӑвашла, ދިވެހިބަސް, буряад, مازِرونی, anarâškielâ, хальмг, հայերեն, West-Vlams, ГӀалгӀай, ဘာသာ မန်, Zazaki,
occitan, მარგალური, Avañe'ẽ, čeština, Tok Pisin, Sesotho sa Leboa, Kiswahili, Boarisch, ქართული, Seeltersk, Latina, Ido, ไทย, मराठी,
norsk, эрзянь, srpskohrvatski / српскохрватски, Lingua Franca Nova, Аҧсшәа, ತುಳು, Taqbaylit, Winaray, Ripoarisch, rumantsch,
मैथिली, русиньскый, kernowek, Chi-Chewa, лакку, Sakizaya, magyar, нохчийн, اردو, Volapük, Deutsch, Ligure, संस्कृतम्, лезги,
asturianu, தமிழ், नेपाली, euskara, سرائیکی, Interlingue, Ladino, Napulitano, डोटेली, қазақша, Malagasy, Gàidhlig, lietuvių, Cymraeg,
Qaraqalpaqsha, हिन्दी, перем коми, فارسی, ślůnski, Türkmençe, slovenčina, Nederlands, isiZulu, Lëtzebuergesch, corsu,
татарча/tatarça, Cebuano, íslenska, монгол, Basa Bali, Acèh, বিষ্ণুপ্রিয়া মণিপুরী, Minangkabau, ئۇيغۇرچە / Uyghurche, Luganda,
chiShona, estremeñu, furlan, Setswana, پښتو, Malti, Nordfriisk, עברית, Kabıyɛ, hornjoserbsce, shqip, Diné bizaad, Pangasinan, polski,
македонски, [illegible], Tiếng Việt, Արեւմտահայերէն, ಕನ್ನಡ, गोंयची कोंकणी / Gõychi Konknni, Ирон, lingála, тоҷикӣ, አማርኛ,
Bahasa Hulontalo, українська, Gaeilge, नेपाल भाषा, བོད་ཡིག, français, Bikol Central, isiXhosa, Esperanto, العربية

Language ID

Language Identification

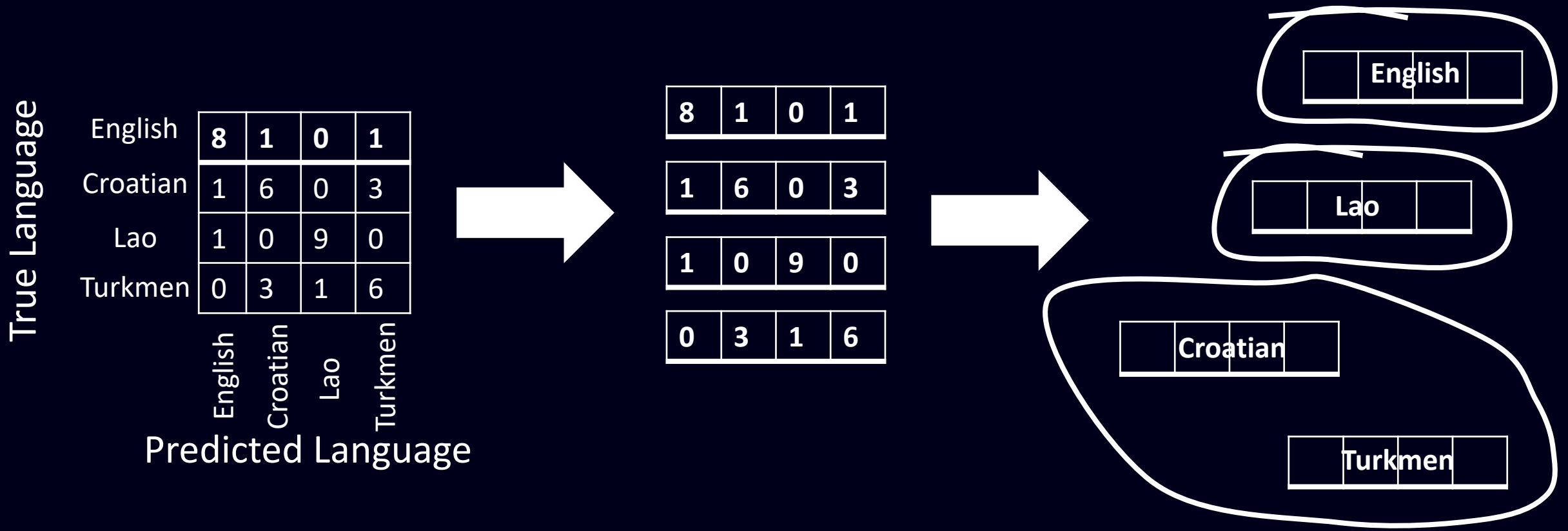
- Count Vectorizer on character bigrams
- Naïve Bayes
- 90.6% accuracy



Relatedness mapping

Mapping Relatedness

- *k*-means clustering, *k=150*



Clusters

- Swahili, Tsonga, Xhosa, Chewa, Kinyarwanda (Bantu)
- French, Norman, Picard (Oïl)
- Persian, Gilaki, Mazanderani (Western Iranian)
- Central Bikol, Cebuano, Javanese, Pangasinan, Tagalog (Malayo-Polynesian), Tok Pisin (English creole)
- Alemannic, Ripuarian, Pennsylvania German, Palatine German (High German)
- Romanian (Balkan Romance), Silesian (West Slavic), Kotava, Lojban (constructed)

Anonymizing writing systems

- For all chunked text in the language, rank all characters (except space) by commonality
- Replace space with 0, most common character with 1, second most with 2, etc up to 256
- Count Vectorizer on character bigrams
- Naïve Bayes
- 99% accuracy

con esta renas la imperos brites

\x0c\x06\x05**\x00**\x02\x03\t\x01**\x00**\x08\x02\x05\x01\x03**\x00**\x07\x01**\x00**\x04

\x0e\r\x02\x08\x06\x03**\x00**\x12\x08\x04\t\x02\x03

Conclusions and future work

- Geographical > Ancestral (eg Romanian, Tok Pisin)
- Writing systems aren’t helpful, or rarer characters aren’t helpful

Future work

- Understand anonymized writing system results
- Explore hyperparameters