Washington State University - Vancouver

Assignment 4

Chelsea Sutcliffe

October 25, 2020

1 Database Schema

```
schema = Schema(id=ID(stored=True), title=TEXT(stored=True),
    description=TEXT(stored=True))
```

The schema stores the tokenid of each dictionary in the index, mainly for logging purposes. Title is the title of the wikipedia page extracted and is stored to display the title back to the user. Description is the field being searched by whoosh and used to return results from the users query.

2 Size of Database

2410 Tuples - Language_whoosh.py prints the number of documents in the index while the server is running.
Database is found in index/languages_spoken.csv.

3 Search Results Returned

3.1 Moroccan Arabic

Moroccan Arabic

Moroccan Arabic (known as Darija in Morocco) is a form of vernacular Arabic spoken in Morocco. It is part of the Maghrebi Arabic dialect continuum, and as such is mutually intelligible to some extent with Algerian Arabic and to a lesser extent with Tunisian

Arabic. It has been heavily influenced mainly by the Berber languages and to a lesser extent by Latin (African Romance), Punic, Persian, French, and Spanish. While Modern Standard Arabic is rarely spoken in daily life and is used to varying degrees in formal situations such as religious sermons, books, newspapers, government communications, news broadcasts or political talkshows, Moroccan Arabic is the spoken common language of Morocco, and has a strong presence in Moroccan television entertainment, cinema and commercial advertising. Sahrawi Hassaniya Arabic spoken in the disputed Moroccan-occupied Western Sahara is usually considered as a separate spoken Arabic variety. Moroccan Arabic has many regional dialects and accents as well. Its mainstream dialect is the one used in Casablanca, Rabat and Fez and therefore it dominates the media eclipsing the other regional dialects like the ones spoken in Tangiers and Oujda. It is spoken as a first language by about 50% to 75% of Morocco's population. The other half speaks one of the Tamazight languages. Educated Moroccan Tamazight-speakers can communicate in mainstream Moroccan Arabic.

Languages of Morocco

There are a number of languages of Morocco. The two official languages are Standard Arabic and Tamazight. Moroccan Arabic (known as Darija) is the spoken native vernacular. The languages of prestige in Morocco are Arabic in its Classical and Modern Standard Forms and some times French, the latter of which serves as a second language for approximately 33% of Moroccans. According to a 2000–2002 survey done by Moha Ennaji, author of Multilingualism, Cultural Identity, and Education in Morocco, "there is a general agreement that Standard Arabic, Moroccan Arabic, and Berber are the national languages." Ennaji also concluded "This survey confirms the idea that multilingualism in Morocco is a vivid sociolinguistic phenomenon, which is favored by many people."There are around 8 million Berber speakers in Morocco. French retains a major place in Morocco, as it is taught universally and serves as Morocco's primary language of commerce and economics, culture, sciences and medicine; it is also widely used in education and government. Morocco is a member of the Francophonie.Spanish is spoken by many Moroccans, particularly in the northern regions around Tetouan and Tangier, as well as in parts of the south, due to historic ties and business interactions with Spain.According to a 2012 study by the Government of Spain, 98% of Moroccans spoke Moroccan Arabic, 63% spoke French, 43% Amazigh, 14% spoke English, and 10% spoke Spanish.

Maghrebi Arabic

Maghrebi Arabic (Western Arabic; as opposed to Eastern or Mashriqi Arabic) is a vernacular Arabic dialect continuum spoken in the Maghreb region, in Morocco, Algeria, Tunisia, Libya, Western Sahara, and Mauritania. It includes Moroccan, Algerian, Tunisian, Libyan, and Hassaniya Arabic. Speakers of Maghrebi Arabic are primarily Arab-Berbers who call their language Derdja, Derja, Derija or Darija. This serves to differentiate the spoken vernacular from Standard Arabic. The Maltese language is believed to be derived from Siculo-Arabic and ultimately from Tunisian Arabic, as it contains some typical Maghrebi Arabic areal characteristics.

Algerian Arabic

Algerian Arabic (known as Darja in Algeria) is a dialect derived from the form of Ara-

bic spoken in northern Algeria. It belongs to the Maghrebi Arabic language continuum and is partially mutually intelligible with Tunisian and Moroccan. Like other varieties of Maghrebi Arabic, Algerian has a mostly Semitic vocabulary. It contains Berber and Latin (African Romance) influences and has numerous loanwords from French, Andalusian Arabic, Ottoman Turkish and Spanish. Algerian Arabic is the native dialect of 75% to 80% of Algerians and is mastered by 85% to 100% of them. It is a spoken language used in daily communication and entertainment, while Modern Standard Arabic (MSA) is generally reserved for official use and education.

Languages of Gibraltar

As Gibraltar is a British overseas territory, its sole official language is English, which is used by the Government and in schools. The eponymous Gibraltarian English accent is spoken in the territory. Most locals are bilingual, also speaking Spanish, because of Gibraltar's proximity to Spain. Most Gibraltarians converse in Llanito, their vernacular which is mostly based on Andalusian Spanish but with numerous loanwords from English as well other Mediterranean languages. However, because of the varied mix of ethnic groups which reside there, other languages such as Moroccan Berber, Moroccan Arabic and Hindi are also spoken on The Rock.

Judeo-Berber language

Judeo-Berber (Berber languages: is any of several hybrid Berber varieties traditionally spoken as a second language in Berber Jewish communities of central and southern Morocco, and perhaps earlier in Algeria. Judeo-Berber is (or was) a contact language; the first language of speakers was Judeo-Arabic. (There were also Jews who spoke Berber as their first language, but not a distinct Jewish variety.) Speakers emigrated to Israel in the 1950s and 1960s. While mutually comprehensible with the Tamazight spoken by most inhabitants of the area (Galand-Pernet et al. 1970:14), these varieties are distinguished by the use of Hebrew loanwords and the pronunciation of š as s (as in many Jewish Moroccan Arabic dialects).

Northern Berber languages

The Northern Berber languages are a dialect continuum spoken across the Maghreb, constituting a subgroup of the Berber branch of the Afroasiatic family. Their continuity has been broken by the spread of Arabic, and to a lesser extent by the Zenati group of Northern Berber. The Zenati idioms share certain innovations not found in the surrounding languages; notably a softening of k to sh and an absence of a- in certain words, such as "hand" (afus vs. fus.) Northern Berber languages spoken by over a million people include Shilha, Central Morocco Tamazight, Riff, Shawiya and Kabyle. They fall into three groups: Moroccan Atlas languages (incl. Shilha, Central Morocco Tamazight) Zenati languages (incl. Riff, Shawiya) KabyleThe eastern boundaries of the North Berber varieties are uncertain. Some linguists include the Nafusi and Ghadames languages, while others do not. Most regard Ghadamès as lying outside of Northern Berber, but the Ethnologue does not. There is no authoritative answer as to whether the Northern Berber varieties constitute languages rather than dialects. Some academics believe that not only Northern Berber but all the Berber languages are dialects of a single language, whereas others come up with much higher counts. At any rate, mutual comprehensibility among the Northern Berber varieties is high, though not perfect.

Berber languages

The Berber languages, also known as Berber or the Amazigh languages, are a branch of the Afroasiatic language family. They comprise a group of closely related languages spoken by the Berbers, who are indigenous to North Africa. The languages were traditionally written with the ancient Libyco-Berber script, which now exists in the form of Tifinagh.Berber is spoken by large populations of Morocco, Algeria and Libya, by smaller populations of Tunisia, northern Mali, western and northern Niger, northern Burkina Faso and Mauritania and in the Siwa Oasis of Egypt. Large Berber-speaking migrant communities, today numbering about 4 million, have been living in Western Europe, spanning over three generations, since the 1950s. The number of Berber people is much higher than the number of Berber speakers. Around 95% of the Berber-speaking population speak one of seven major varieties of Berber, each with at least 2 million speakers. The now extinct Guanche language spoken on the Canary Islands by the Guanches, as well as possibly the languages of the ancient C-Group culture in today's southern Egypt and northern Sudan, are believed to have belonged to the Berber branch of the Afroasiatic family. The Berber languages and dialects have had a written tradition, on and off, for about 2,500 years, although the tradition has been frequently disrupted by cultural shifts and invasions. They were first written in the Libyco-Berber abjad, which is still used today by the Tuareg in the form of Tifinagh. The oldest dated inscription is from the 3rd century BCE. Later, between about 1000 CE and 1500 CE, they were written in the Arabic script, and since the 20th century they have been written in the Berber Latin alphabet, especially among the Kabyle and Riffian communities of Morocco and Algeria. The Berber Latin alphabet was also used by most European and Berber linguists during the 19th and 20th centuries.A modernised form of the Tifinagh alphabet, called Neo-Tifinagh, was adopted in Morocco in 2003 for writing Berber, but many Moroccan Berber publications still use the Berber Latin alphabet. Algerians mostly use the Berber Latin alphabet in Berber-language education at public schools, while Tifinagh is mostly used for artistic symbolism. Mali and Niger recognise a Tuareg Berber Latin alphabet customised to the Tuareg phonological system. However, traditional Tifinagh is still used in those countries. There is a cultural and political movement among speakers of the closely related varieties of Northern Berber to promote and unify them under a written standard language called Tamazigt (or Tamazight). The name Tamazigt is the current native name of the Berber language in the Moroccan Middle Atlas and Rif regions and the Libyan Zuwarah region. In other Berber-speaking areas, this name was lost. There is historical evidence from medieval Berber manuscripts that all indigenous North Africans from Libya to Morocco have at some point called their language Tamazigt. The name Tamazigt is currently being used increasingly by educated Berbers to refer to the written Berber language, and even to Berber as a whole, including Tuareg. In 2001, Berber became a constitutional national language of Algeria, and in 2011 Berber became a constitutionally official language of Morocco. In 2016, Berber became a constitutionally official language of Algeria alongside Arabic.

3.2 Romance Latin

Vulgar Latin

Vulgar Latin or Sermo Vulgaris ("common speech"), also Colloquial Latin, or Common Romance (particularly in the late stage), was a range of non-standard sociolects of Latin spoken in the Mediterranean region during and after the classical period of the Roman Empire. It is distinct from Classical Latin, the standard and literary version of the language. Compared to Classical Latin, written documentation of Vulgar Latin appears less standardized. Works written in Latin during classical times and the earlier Middle Ages used prescribed Classical Latin rather than Vulgar Latin, with very few exceptions (most notably sections of Gaius Petronius' Satyricon), thus Vulgar Latin had no official orthography of its own. By its nature, Vulgar Latin varied greatly by region and by time period, though several major divisions can be seen. Vulgar Latin dialects began to significantly diverge from Classical Latin by the third century during the classical period of the Roman Empire. Nevertheless, throughout the sixth century, the most widely spoken dialects were still similar to and mostly mutually intelligible with Classical Latin. In terms of regional differences for the whole Latin period, "we can only glimpse a tiny amount of divergence with the actual written data. In texts of all kinds, literary, technical, and all others, the written Latin of the first five or six centuries A.D. looks as if it were territorially homogeneous, even in its 'vulgar' register. It is only in the later texts, of the seventh and eighth centuries, that we are able to see in the texts geographical differences that seem to be the precursors of similar differences in the subsequent Romance languages."In the Eastern Roman Empire, Latin gradually faded as the Court language over the course of the 6th century (official Latin lost its predominance in official communications from the mid-5th, although all communications at the imperial level of administration in Greek had to be accompanied by a Latin text); it was used in Justinian's (whose native language was Latin), but during the reign of Heraclius in the early 7th century, Greek (which was already widely spoken in the eastern portions of the Roman Empire from its inception) was made the official language. The Vulgar Latin spoken in the Balkans north of Greece became heavily influenced by Greek and Slavic (Vulgar Latin already had Greek loanwords before the Roman Empire) and also became radically different from Classical Latin and from the proto-Romance of Western Europe.Thus the Latin of classical antiquity changed from being a "living natural mother tongue" to being a language foreign to all, which could not be used or understood even by Romance-speakers except as a result of deliberate and systematic study. If a date is wanted "we could say Latin 'died' in the first part of the eighth century", and after a long period 650–800 A.D. of rapidly accelerating changes. Even after the end of Classical Latin, people had no other names for the languages they spoke than Latin, lingua romana, or lingua romana rustica (to distinguish it from formal Latin) for 200–300 years. The Romance languages, such as Catalan, French, Italian, Occitan, Portuguese, Romanian, and Spanish all evolved from Vulgar Latin and not from Classical Latin, but linguists prefer to distinguish the attested Vulgar Latin from the reconstructed model of Proto-Romance.

Romance-speaking Africa

Romance-speaking Africa or Latin Africa consists of the countries and territories in Africa whose official or main languages are Romance ones, and countries which have significant populations that speak Romance languages: French, Portuguese, Spanish, and Italian. Many of these countries are members of the Organisation internationale de la Francophonie (OIF; International Organization of La Francophonie) or the Community of Portuguese Language Countries (Comunidade dos Países de Língua Portuguesa), and seven are members of the Latin Union.North Africa, from Morocco to Egypt, was part of the Roman Empire. As a result, the African Romance language evolved in Tunisia, Algeria and Morocco. It was spoken until the 13th century.

Roman language

Roman language may refer to: Latin, the language of Ancient Rome Languages of the Roman Empire Romance languages, the languages descended from Latin, including French, Spanish and Italian Romanesco dialect, the variety of Italian spoken in the area of Rome

Romance languages

The Romance languages (less commonly Latin languages, or Neo-Latin languages) are the modern languages that evolved from Vulgar Latin between the third and eighth centuries. They are a subgroup of the Italic languages in the Indo-European language family. The five most widely spoken Romance languages by number of native speakers are Spanish (480 million), Portuguese (255 million), French (77 million), Italian (65 million), and Romanian (24 million). Among the various Romance languages, Sardinian and Italian are the closest to Latin, followed by Spanish, Romanian, Portuguese, and the most divergent being French.The more than 900 million native speakers of Romance languages are found worldwide, mainly in the Americas, Europe, and parts of Africa, as well as elsewhere. The major Romance languages also have many non-native speakers, and are in widespread use as lingua francas. This is especially true of French, which is in widespread use throughout Central and West Africa, Madagascar, Mauritius, Seychelles, Comoros, Djibouti, Lebanon, and North Africa (excluding Egypt, where it is a minority language). Because it is difficult to assign rigid categories to phenomena such as languages which exist on a continuum, estimates of the number of modern Romance languages vary. For example, Dalby lists 23, based on the criterion of mutual intelligibility. The following includes those and additional current, living languages, and one extinct language, Dalmatian: Ibero-Romance: Portuguese, Galician, Mirandese, Asturian, Leonese, Spanish, Aragonese, Ladino (Judaeo-Spanish); Occitano-Romance: Catalan/Valencian, Occitan (lenga d'oc), Gascon; Gallo-Romance: French/Oïl languages, Franco-Provençal (Arpitan); Rhaeto-Romance: Romansh, Ladin, Friulian; Gallo-Italic: Piedmontese, Ligurian, Lombard, Emilian-Romagnol; Italo-Dalmatian: Italian, Tuscan, Romanesco, Corsican, Sassarese, Sicilian, Neapolitan, Dalmatian (extinct in 1898), Venetian (classification disputed), Istriot; Sardinian; Eastern Romance: Romanian (standard known as Daco-Romanian), Istro-Romanian, Aromanian, Megleno-Romanian.

Spanish language

Spanish or Castilian is a Romance language that originated in the Iberian Peninsula of Europe and today is a global language with more than 483 million native speakers, mainly in Spain and the Americas. It is the world's second-most spoken native

language, after Mandarin Chinese, and the world's fourth-most spoken language, after English, Mandarin Chinese and Hindi. Spanish is a part of the Ibero-Romance group of languages, which evolved from several dialects of Vulgar Latin in Iberia after the collapse of the Western Roman Empire in the 5th century. The oldest Latin texts with traces of Spanish come from mid-northern Iberia in the 9th century, and the first systematic written use of the language happened in Toledo, a prominent city of the Kingdom of Castile, in the 13th century. Beginning in 1492, the Spanish language was taken to the viceroyalties of the Spanish Empire, most notably to the Americas, as well as territories in Africa, Oceania and the Philippines.A 1949 study by Italian-American linguist Mario Pei, analyzing the degree of difference from a language's parent (Latin, in the case of Romance languages) by comparing phonology, inflection, syntax, vocabulary, and intonation, indicated the following percentages (the higher the percentage, the greater the distance from Latin): In the case of Spanish, it is one of the closest Romance languages to Latin (20% distance), only behind Sardinian (8% distance) and Italian (12% distance). Around 75% of modern Spanish vocabulary is derived from Latin, including Latin borrowings from Ancient Greek. Spanish vocabulary has been in contact with Arabic from an early date, having developed during the Al-Andalus era in the Iberian Peninsula and around 8% of its vocabulary has an Arabic lexical root. It has also been influenced by Basque, Iberian, Celtiberian, Visigothic, and other neighboring Ibero-Romance languages. Additionally, it has absorbed vocabulary from other languages, particularly other Romance languages—French, Italian, Andalusi Romance, Portuguese, Galician, Catalan, Occitan, and Sardinian—as well as from Quechua, Nahuatl, and other indigenous languages of the Americas.Spanish is one of the six official languages of the United Nations. It is also used as an official language by the European Union, the Organization of American States, the Union of South American Nations, the Community of Latin American and Caribbean States, the African Union and many other international organizations.Despite its large number of speakers, the Spanish language does not feature prominently in scientific writing, though it is better represented in the humanities. Approximately 75% of scientific production in Spanish is divided into three thematic areas: social sciences, medical sciences and arts/humanities. Spanish is the third most used language on the internet after English and Chinese.

Latin

Latin is a classical language belonging to the Italic branch of the Indo-European languages. Latin was originally spoken in the area around Rome, known as Latium. Through the power of the Roman Republic, it became the dominant language in Italy, and subsequently throughout the western Roman Empire. Latin has contributed many words to the English language. In particular, Latin (and Ancient Greek) roots are used in English descriptions of theology, the sciences, medicine, and law. It is the official language in the Holy See (Vatican City). By the late Roman Republic (75 BC), Old Latin had been standardised into Classical Latin. Vulgar Latin was the colloquial form spoken during the same time and attested in inscriptions and the works of comic playwrights like Plautus and Terence and author Petronius. Late Latin is the written language from the 3rd century; its colloquial form Vulgar Latin developed in the 6th to 9th centuries into the Romance languages, such as Italian, Sardinian, Venetian, Neapolitan, Sicilian,

Piedmontese, Lombard, French, Franco-Provençal, Occitan, Corsican, Ladin, Friulan, Romansh, Catalan/Valencian, Aragonese, Spanish, Asturian, Galician, and Portuguese. Medieval Latin was used as a literary language from the 9th century to the Renaissance which used Renaissance Latin. Later, Early Modern Latin and New Latin evolved. Latin was the language of international communication, scholarship and science until well into the 18th century, when vernaculars (including the Romance languages) supplanted it. Ecclesiastical Latin remains the official language of the Holy See and the Roman Rite of the Catholic Church. Latin is a highly inflected language, with three distinct genders, six or seven noun cases, five declensions, four verb conjugations, six tenses, three persons, three moods, two voices, two or three aspects and two numbers. The Latin alphabet is derived from the Etruscan and Greek alphabets and ultimately from the Phoenician alphabet. Iberian Romance languages

The Iberian Romance, Ibero-Romance or simply Iberian languages, is an areal grouping of Romance languages that developed on the Iberian Peninsula, an area consisting primarily of Spain, Portugal, Gibraltar and Andorra, and in southern France which are today more commonly separated into West Iberian and Occitano-Romance language groups. Evolved from the Vulgar Latin of Iberia, the most widely spoken Iberian Romance languages are Spanish, Portuguese, Catalan-Valencian-Balear and Galician. These languages also have their own regional and local varieties. Based on mutual intelligibility, Dalby counts seven "outer" languages, or language groups: Galician-Portuguese, Spanish, Astur-Leonese, "Wider"-Aragonese, "Wider"-Catalan, Provençal+Lengadocian, and "Wider"-Gascon.In addition to those languages, there are a number of Portuguese-based creole languages and Spanish-based creole languages, for instance Papiamento.

Italic languages

The Italic languages form a branch of the Indo-European language family, whose earliest known members were spoken in the Italian Peninsula in the first millennium BC. The best known of them is Latin, the official language of the Roman Empire, which conquered the other Italic peoples before the common era. The other Italic languages became extinct in the first centuries AD as their speakers were assimilated into the Roman Empire and shifted to some form of Latin. Between the third and eighth centuries AD, Vulgar Latin (perhaps influenced by language-shift from the other Italic languages) diversified into the Romance languages, which are the only Italic languages natively spoken today. Besides Latin, the known ancient Italic languages are Faliscan (the closest to Latin), Umbrian and Oscan (or Osco-Umbrian), and South Picene. Other Indo-European languages once spoken in the peninsula, whose inclusion in the Italic branch is disputed, are Aequian, Vestinian, Venetic and Sicel. These long-extinct languages are known only from inscriptions in archaeological finds. In the first millennium BC, several (other) non-Italic languages were spoken in the peninsula, including members of other branches of Indo-European (such as Celtic and Greek) as well as at least one non-Indo-European one, Etruscan. It is generally believed that those 1st millennium Italic languages descend from Indo-European languages brought by migrants to the peninsula sometime in the 2nd millennium BC. However, the source of those migrations and the history of the languages in the peninsula are still the matter of debate among historians. In particular, it is debated whether the ancient Italic languages all descended from

a single Proto-Italic language after its arrival in the region, or whether the migrants brought two or more Indo-European languages that were only distantly related. With over 800 million native speakers, the Romance languages make Italic the second-most-widely spoken branch of the Indo-European family, after Indo-Iranian. However, in academia the ancient Italic languages form a separate field of study from the medieval and modern Romance languages. This article focuses on the ancient languages. For the others, see Romance studies. All Italic languages (including Romance) are generally written in Old Italic scripts (or the descendant Latin alphabet and its adaptations), which descend from the alphabet used to write the non-Italic Etruscan language, and ultimately from the Greek alphabet.

Foreign language influences in English

The core of the English language descends from Old English, the language brought with the Angle, Saxon, and Jutish settlers to what was to be called England from the 500s. The bulk of the language in spoken and written texts is from this source. As a statistical rule, around 70 percent of words in any text are Anglo-Saxon. Moreover, the grammar is largely Anglo-Saxon.A significant portion of the English vocabulary comes from Romance and Latinate sources. Estimates of native words (derived from Old English) range from 20%–33%, with the rest made up of outside borrowings. A portion of these borrowings come directly from Latin, or through one of the Romance languages, particularly Anglo-Norman and French, but some also from Italian, Portuguese, and Spanish; or from other languages (such as Gothic, Frankish or Greek) into Latin and then into English. The influence of Latin in English, therefore, is primarily lexical in nature, being confined mainly to words derived from Latin roots.While some new words enter English as slang, most do not. Some words are adopted from other languages; some are mixtures of existing words (portmanteau words), and some are new creations made of roots from dead languages.

British Latin

British Latin or British Vulgar Latin was the Vulgar Latin spoken in Great Britain in the Roman and sub-Roman periods. While Britain formed part of the Roman Empire, Latin became the principal language of the elite, especially in the more Romanised south and east of the island. However, in the less Romanised north and west it never substantially replaced the Brittonic language of the indigenous Britons. In recent years, scholars have debated the extent to which British Latin was distinguishable from its continental counterparts, which developed into the Romance languages. After the end of Roman rule, Latin was displaced as a spoken language by Old English in most of what became England during the Anglo-Saxon settlement of the fifth and sixth centuries. It survived in the remaining Celtic regions of western Britain and had died out by about 700, when it was replaced by the local Brittonic languages.