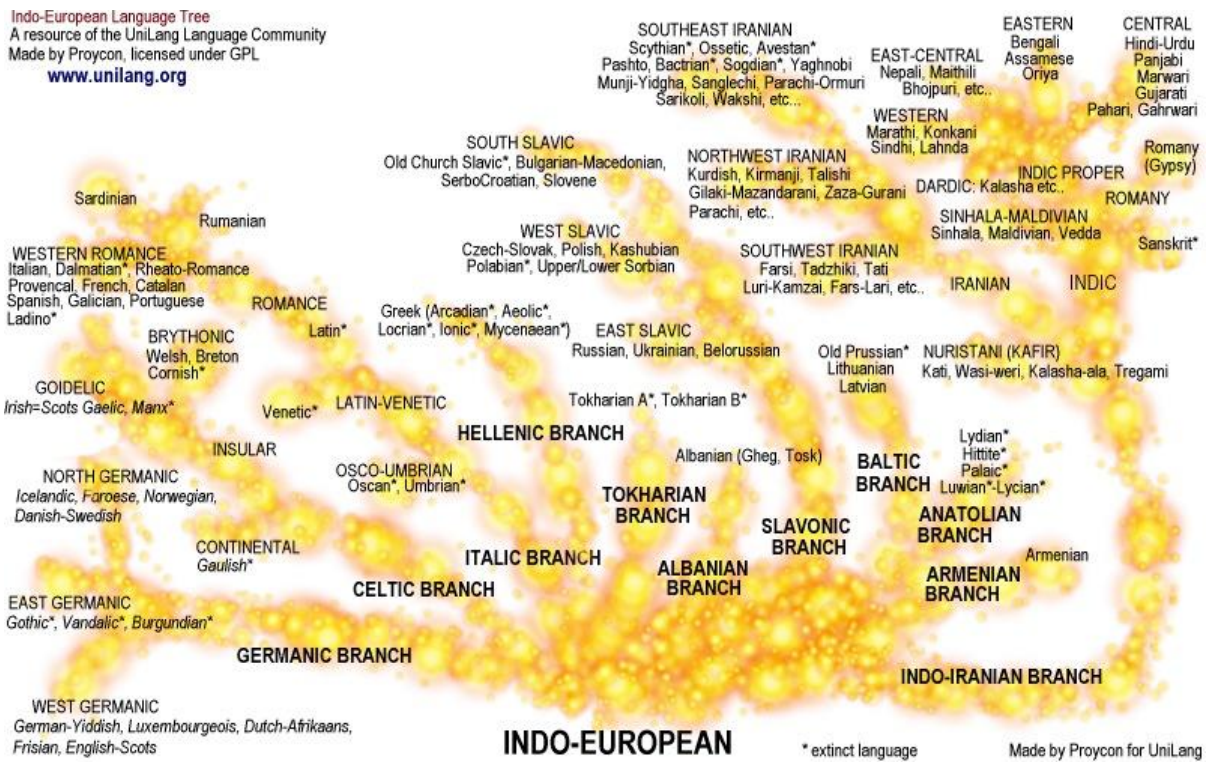
Introduction to Data Mining

2. domača naloga: podobnost jezikov

Slika prikazuje hierarhijo indoevropskih jezikov. Jeziki so postavljeni v hierarhijo glede na izvor, kar lepo sovpada s podobnostjo med njimi. V tej nalogi boste z analizo podobnosti jezike razvrstili v skupine.



Priložena datoteka vsebuje Splošno deklaracijo človekovih pravic v 263 jezikih. To besedilo je prevedeno v rekordno število jezikov: spletna stran Združenih narodov jih ponuja kar 503 (v priloženi datoteki so zgolj nekateri). Deklaracijo v zapisu UTF-8 odprite z `open(datoteka, "rt", encoding="utf8").read()`.

Izmed ponujenih jezikov jih za razvrščanje izberite vsaj dvajset (med njimi naj bodo vsaj štirje slovanski, romanski, in germanski jeziki). Poleg jezikov v latinici vključite še jezike v vsaj dveh drugih pisavah; denimo v cirilici ali v grški pisavi. Problem različnih abeced rešite s transliteracijo s knjižnico unidecode.

1. Implementirajte postopek za razvrščanje v skupine na podlagi medoidov (*k*-medoids clustering), ki razdalje med besedili meri s kosinusno razdaljo s primerjavo frekvenc trojk sosednjih črk in to **nujno** brez uporabe polnih matrik. Pri računanju razdalj morate torej upoštevati le trojke, ki jih dani besedili dejansko vsebujeta; program bi moral delovati podobno hitro, četudi bi namesto trojk primerjali enajsterke. [50%]
2. Razviti postopek poženite s 100 naključno izbranimi inicializacijami oziroma začetnim izborom *k=5* medoidov. Postopek razvrščanja z medoidi je lahko namreč precej odvisen od začnega izbora medoidov. Na ta način boste dobili 100 različnih razvrstitev oziroma rezultatov postopka. Vsakega od 100 razvrstitev ocenite z metodo silhuete, ki jo razvijete sami. Izrišite tudi porazdelitev (histogram) vrednosti silhuet. [30%]
3. Kodo za računanje razdalj nadgradite s funkcijo, ki zna za poljubno besedilo (v enem izmed izbranih jezikov) ugotoviti jezik, v katerem je besedilo napisano. Predlagajte in implementirajte tehniko, ki za dano besedilo oceni, s kakšno verjetnostjo pripada v vašem programu zajetim jezikom. Poročajte npr. o pripadnosti trem najbolj verjetnim jezikom. Funkcijo preizkusite na desetih malo daljših stavkih ali odlomkih, ki jih spišete sami ali pa poiščete kje na internetu. [20%]

Oddaja: Oddajte poročilo in izvorno kodo (z uporabljenimi podatki), oboje v skupni .zip datoteki. Koda naj bo spisana v eni sami datoteki. Poročilo mora biti napisano s predpisano predlogo (LaTeX). Pri pisanju poročila ne uporabljajte razdelkov iz predloge, ampak vključite zgolj naslednje razdelke:

1. **Izbrani jeziki**. Naštejte izbrane jezike ter opišite vašo predobdelavo datotek.
2. **Rezultati razvrščanja.** Vključite porazdelitev vrednosti silhuet ter rezultate razvrščanja (skupine jezikov) pri razvrščanju z najboljšo in najslabšo silhueto. V največ treh stavkih komentirajte smiselnost rezultatov.
3. **Napovedovanje jezika.** V enem odstavku opišite vaš predlagani postopek. V poročilo vključite tabelo s kratkimi besednimi odlomki in napovedi verjetnosti treh najbolj verjetnih razredov.

Dodatno (+10%): Na istih podatkih in z isto tehniko računanja razdalj med besedili uporabite tudi hierarhično razvrščanje. Na kratko (v največ treh stavkih) komentirajte rezultate.

Dodatno (+10%): Na spletu najdite novičarske strani v prej izbranih dvajsetih jezikih in ponovite analizo na novicah. Komentirajte rezultate.

◄ 1. domača naloga: glasovanje za Pesem Evrovizije

Jump to...

3. domača naloga: napovedovanje prihodov avtobusov LPP ►