**Opis problému:**

Odporúčací systém StumbleUpon odporúča svojim užívateľom stránky, pričom si potrpí na vysokej kvalite týchto stránok. Kvalita stránky je aj to, či bude stabilná, alebo dlhodobo úspešná. Takéto „evergreen“ stránky sa bežne zisťujú skrz používateľskú odozvu, avšak problémom je ako tento údaj vedieť ešte predtým, než je stránka vôbec hocikomu odporúčaná. Existuje predpoklad, že určité vlastnosti stránky by mohli indikovať, či má stránka potenciál na to byť „evergreen“. Práve preto je toto úloha pre strojové učenie , konkrétne použitie klasifikátora, ktorou sa budeme zaoberať v našom projekte.

**Opis dát:**

Ako zdrojov dát je dostupných niekoľko súborov vo formáte tsv, angl. „tab-separated values“. Prvý zhluk týchto skomprimovaných súborov predstavujú obsahy jednotlivých stránok zapísané v tomto formáte. Ako názvy súborov sú uvedené ID stránok.

Ďalšie dva tsv súbory sú nazvané train a test, pretože je jeden určený pre trénovanie, obsahujúci 7395 stránok, a druhý pre testovanie, obsahujúci 3171 stránok. Tieto súbory obsahujú meta-údaje o stránkach, obsahujúce niekoľko atribútov:

* **Url stránky** a **ID stránky**, ktoré je rovnaké ako názov súboru obsahujúceho obsah danej stránky v zhluku komprimovaných tsv súborov.
* **Boilerplate**, obsahujúci krátky opis stránky v json formáte.
* **Alchemy kategória a skóre** z verejne dostupného ohodnocovača stránok Alchemy API.
* **Veľkostné údaje**, medzi ktoré patria priemerný počet slov na linku stránky, celkový počet textových slov a počet tagov, ako sú: <embed>,<a>.
* **Pomerové údaje**, medzi ktoré patria podobnosť odkazov počítaná percentom odkazov obsahujúcich 1,2,3, alebo 4 podobné slová, dosiahnutá kompresia (pomer redundantných informácií), pomer <iframe> tagov voči celkovému počtu tagov, pomer tagov voči textu, pomer <img> tagov voči textu, percento slov na stránke, ktoré nie sú súčasťou linku a nakoniec pomer slov nenájdených vo wiki (chybne napísaných).
* **Príznakové údaje**, medzi ktoré patria príznak, či má stránka <frame> tag, ale nemá žiadny <body> tag, príznak, či stránka obsahuje doménový odkaz s tagom <a>, príznak, či je stránka klasifikovaná StumbleUpon klasifikátorom ako správa, alebo správa na titulnej strane, príznak, že aspoň 3 odkazy obsahujú viac ako 30 alfanumerických znakov a nakoniec pre trénovaciu množinu príznak, či je stránka „evergreen“, alebo nie.

**Stručný opis prác iných autorov**

<http://dl.acm.org/citation.cfm?id=360406>

Všeobecné informácie o klasifikátoroch a ich vhodnosti použitia od typu úlohy.

<http://dl.acm.org/citation.cfm?id=1459357>

http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=4119127

Metódy klasifikovania web stránok z textu a metadát.

<http://dl.acm.org/citation.cfm?id=846188>

Web mining vo všeobecnosti.

**Predspracovanie atribútov**

Pre dané atribúty sme spravili štatistickú analýzu, na základe ktorej sa odfiltruje podmnožina dimenzií, ktoré budú predstavovať vstupné údaje do klasifikátora. Pri filtrácii sme zohľadnili chýbajúce údaje, logickú použiteľnosť každého atribútu pre klasifikačnú úlohu a korelačné vzťahy každého s každým atribútom. Výsledné vyradené atribúty sú:

* **Embed Ratio** atribút obsahuje hodnoty, ako výsledky pomeru, so silným zastúpením hodnôt -1 a 0 (6900 hodnôt z 7395), pričom zvyšné hodnoty sú v rozsahu 0-0.2. Na koľko nepoznáme interpretáciu týchto hodnôt a zdá sa nám, že pomer zmysluplných hodnôt je príliš malý, je tiež tento atribút vyradený.
* **Frame Based** nezohľadňujeme kvôli tomu, že obsahuje samé nuly a je to pokazený atribút.
* Zvyšné 3 tribúty **commonlink\_ratio** boli odstránenéna základe ich silnej korelácie.

Ostatné atribúty vidíme ako potencionálne použiteľné pre klasifikátor.

Zostávajúce atribúty však bolo nutné predspracovať a to doplnením chýbajúcich údajov, odstránením odľahlých hodnôt a normalizáciou.

Prvým a aj najpracnejším atribútom pre predspracovanie je textový obsah webových stránok – tzv. surové dáta (raw data). Tieto dáta sú uložené v JSON (JavaScript Object Notation) formáte. Rozdelené sú na tri časti: titul (title), body (textový obsah stránky) a url stránky. Práve samotný textový obsah je z týchto najzaujímavejší.

Textový obsah sa najprv normalizuje. V rámci normalizácie sa vykonajú nasledovné operácie:

* Transformácia textu do malých písmen (lowercase).
* Odstránenie všetkých ne-písmenkových znakov.

Táto forma normalizácie je vhodná na analýzu obsahu textu, napr. kľúčové slová, alebo tému textu.

Základnou formou pri väčšine techník na spracovanie textových dát je dátová štruktúra vrece slov (bag of words). Takéto vrece obsahuje všetky unikátne slová zo vstupného dokumentu. Okrem samotného vreca slov, počítame aj frekvenciu výskytu jednotlivých slov v dokumente a lexikálnu rôznorodosť dokumentu – pomer medzi počtom unikátnych slov a celkovým počtom slov v dokumente.

V rámci spracovania textového obsahu sme sa zamerali na určovanie témy, resp. kategórie, či žánru textu, napr. či je text o zdraviu, politiky alebo športe a podobne. Tejto problematike sa hovorí klasifikácia dokumentu. Dva základné prístupy ku klasifikácii dokumentov sú klasifikácia s učiteľom a bez učiteľa (supervised and unsupervised classification).

Hlavným problémom klasifikácie s učiteľom je nedostatok vhodných voľne dostupných oanotovaných textových datasetov. Z tohto dôvodu sme sa rozhodli využiť dáta, ktoré máme k dispozícii. Zhruba 70% záznamov majú určenú kategóriu textového obsahu – atribút alchemy category. Natrénovaním klasifikátora na týchto dátach sa môžu jednotlivé kategórie priradiť všetkých textom. Na tento účel sme sa rozhodli použiť najpoužívanejší modul na analýzu textu NLTK (Natural Language Toolkit) pre Python. Tento ponúka naivný Baiesov klasifikátor a rozhodujúce stromy. Oba si však vyžadujú ukladanie slov do pythonovkých slovníkov, ktoré pri našich textových dátach dosiahli veľkú pamäťovú náročnosť – na notebooku s veľkosťou RAM pamäte 12 GB nám zmrzol celý operačný systém po asi desiatich minútach behu skriptu.

Keďže je analýza textových dát sama o sebe osobitným problémom a nedopracovali sme sa k žiadnym alternatívam, rozhodli sme sa využiť iba jednoduché štatistické atribúty pri analýze textových dát.

*Poznámka:* Klasifikáciu bez učiteľa sme zatiaľ nestihli vyskúšať a je možné, že sa nám s touto bude lepšie dariť.

V rámci predspracovania zvyšných atribútov sa:

* doplnili nuly do chýbajúcich hodnôt atribútu **News Front Page**, nakoľko je prehnaná väčšina hodnôt 0,
* odstránili odľahlé hodnoty v atribúte **Image ratio** okresaním hodnôt na interval 0-1, pričom za chybné hodnoty boli doplnené priemerné hodnoty
* vytvorila funkcia na normalizáciu atribútov, ktorá sa využije v prípade potreby (najmä pri klasifikátore neurónovej siete)

**DM metódy**

Našim výstupom je model klasifikátora, ktorý je schopný určiť „evergreen“ vlastnosť stránky s čo najvyššou úspešnosťou, pričom sa takto tiež získajú znalosti o tom, ktoré atribúty webových stránok najviac na túto vlastnosť vplývajú.

Na túto úlohu je vhodných viacero „supervised“ klasifikátorov, ktorými sú:

* Rozhodovacie stromy – Strom pravidiel, pričom slabšia presnosť, avšak vyššia rýchlosť.
* Support vector machines – Nová, úspešná metóda. Problém s prevedením nelineárnych dát.
* Neurónové siete – Nie je vidno „do vnútra“ výpočtov siete, avšak vysoká úspešnosť. Nutnosť normalizácie atribútov
* Bayersovské siete – Siete rovníc pravdepodobností, úspešnosť podobná neurónovým sieťám.

**Prvotné experimenty**

**-**dodáme do odovzdania**-**

**Vyhodnocovanie**

Ako vyhodnocovaciu mieru budeme používať chybu klasifikácie, počítanú ako MAPE, alebo MSE, čo je suma štvorcov rozdielov medzi klasifikovanou a korektnou hodnotou. Takisto môžeme použiť ROC krivku vzhľadom na mieru presnosti a špecifickosti klasifikácie.