Gəlin görək, tək sənədlərdən bu tip mətn parçalarını

və ya digər xülasələri necə yaratmaq olar.

Bu mətn parçası nümunəsidir, yenə də Google axtarış sistemin

gəlir. Mən bele bir sual vermişəm: "was cast metal

movable type invented in Korea?" Və Google mənə cavabla birlikdə

üç balaca mətn parçası verir və görə bilərsiniz ki, həmin mətn

parçasındakı bəzi sözlər qalın şəkildə göstərilib, çünki həmin

sözlər bizim göndərdiyimiz sorğudakı sözlərlə eynidir.

Və görə bilərsiniz ki, mətn parçasındakı üç nöqtə onu

göstərir ki, sistem həmin mətn parçasının hissələrini səhifənin

müxtəlif yerlərindən götürür. Gəlin bu tip tək sənədə əsaslanan

mətn parçaları və digər növ qısa məzmunlar necə yaradılır. Siz

istənilən qısa məzmun çıxartma alqoritminin üç mərhələdən ibarət

olduğunu düşünə bilərsiniz.

Birinci mərhələ məzmun seçmədir (content selection):

sənəddən bizə lazım olan cümlələrin seçilməsi. Məsəl üçün bizim bir

sənədimiz var və bizim həmin sənəddən cümlələri çıxartmağa

ehtiyacımız var. Biz cümlələr toplusunu müxtəlif cümlələrə

parçalamalıyıq. Ola bilsin biz tam cümlələri seçmək istəyirik.

Və ya biz bir növ hərəkət edən pəncərədən istifadə edə bilərik.

Beləliklə, biz kiçik hissələri, kiçik cümlələri çıxardıq və bu

seçilmiş cümlələr qrupundan bizim üçün önəmli olanları götürmək

istəyirik və həmin cümlələri mən kiçik qara nöqtələrlə qeyd etmişəm.

Biz cümlələri götürdükdən sonra

informasiya sıralanması (information ordering) adlı

növbəti mərhələdə cümlələri necə sıralayacağımıza baxacağıq.

Bundan sonra artıq əlimizdə sıralanmış önəmli cümlələr var və

həmin cümlələrə bəzi dəyişiklər tətbiq edə bilərik: biz

onları sadələşdirə və ya başqa sey edə bilərik. Bu cümlə reallışdırma

mərhələsi (sentence realization) adlanır və yerinə yetirdiyimiz

üç mərhələnin nəticəsi isə xülasədir.

Ən sadə qısa məzmun yaratma alqorimti, yəni çox istifadə olunan

birisi ancaq bu mərhələrin birini yerinə yetirir ki, bu da

məzmun seçmə mərhələsidir. Mümkün ən sadə alqoritmdə biz

cümlələrin sıralanması ilə maraqlanmırıq və cümlələrdə

heç bir dəyişiklik etmirik. Biz sadəcə sənədimizi cümlə

seqmentlərinə və ya pəncərələrə bölürük. Önəmli olanları

götürürük və onlar sənəddə hansı sırada idisə elə də

saxalayırıq. Yəni orijinal cümlələr üçün sadəcə sənəddəki

sıranı götürük. Və bu qısa məzmun çıxartmada əsasdır (ana xətt) və

çoxlu veb-əsaslı mətn parçası yaradan alqoritmlərin işini təşkil

edir. Məzmun seçilməsi üçün ən çox istifadə olunan alqoritm

1958-ci yazılmış məqaladə qeyd olunmuşdur və bu tipli ideyaların

bele tez ortaya çıxması həyəcan vericidir. Və intuisiya çox sadədir.

Sadəcə içərisində səciyyəvi və ya informativ sözlər olan

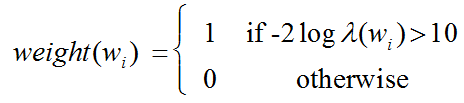
cümlələri seçirik. Bəs bu nə deməkdir? Siz TFİDF-i görürsüz.

* Səciyyəvi sözləri təyin etmək üçün iki yanaşma var:
  1. tf-idf: tf-idf'lə *j* sənədində hər wi sözü üçün çəki əmsalı ver.

weight.PNG

2. mövzu imzası: səciyyəvi sözlərin daha kiçik dəstini seçmək

* + - qarşılıqlı informasiya
    - loqarifma ehtimal nisbəti (LEN) Dunning (1993), Lin and Hovy (2000)



Bu (tf-idf) tez-tez təkarlanan sözlərin seçilmə yoludur və bundan sonra isə bütün sənədlərdə

təkrarlanmayan sözləri seçirik. Yəni bu səciyyəviliyi və ya informativliyi təyin etmək

üçün olan metodlardan biridir. Qısa məzmun yaratmada biz başqa üsuldan da istifadə edə

bilərik. Bu loqarifma ehtimal nisbəti və ya mövzu imzası yanaşması adlanır. Və bu TFİDF-dən

iki yolla fərqlənir. Birincisi biz seçim etmək üçün, yəni, hər bir sözə çəki əmsalı vermək üçün

bir az fərqli statistikadan istifadə edirik. Və ikincisi bütün sözləri götürməkdənsə, biz ancaq

çəki əmsalı müəyyən limitdən böyük olan sözləri götürük, hansı ki, bu sözlər çox səciyyəvidir.

Loqarifma ehtimal nisbəti bizə **lambda** adlanan statistik kəmiyyət verir. Spesifik detallardan

danışmayacam, ancaq məqalələrdən daha çox məlumat əldə etmək mümkündür. Biz mənfi

ikinin lambdanın onluq loqarifması ilə hasilinin 10-dan böyük halı üçün sözləri götürəcəyik. Bu

bizə xüsusi səciyyəvi sözləri seçmək limit verəcək. Beləliklə, biz hər bir sözə çəki əmsalı

verəcəyik. Və əgər söz sənədlə xüsusilə əlaqəlidirsə, yəni verilən limitdən çox təkrarlanırsa

sözün çəki əmsalı 1 olacaq. Digər halda biz sözə çəki əmsalı olaraq 0 veririk. Logafirma

ehtimal nisbəti necə hesablanması və statistika haqqında intuisiya barədə əlavə məlumatı

Ted Dunning'in məqaləsindən və ya Lin və Hovy'nin məqaləsindən (LEN-in qısa məzmun

çıxartmaya tətbiqi) əldə edə bilərsiz. Biz indi sorğu-əsaslı qısa məzmun çıxartmanı yerinə

yetirmək üçün bu alqoritmdə bəzi dəyişiklər etmək istəyirik. Biz bugünki mühazirədə şəffaf

qısa məzmun çıxartma ilə o qədər maraqlanmırıq, biz ancaq suala cavab vermək

üçün qısa məzmun çıxartma metodları ilə maraqlanırıq. Belə ki, bu mövzu imzası əsaslıdır.

Mövzu imzası o deməkdir ki, sənədlə xüsusilə əlaqəli olan sözləri seç. Məzmun seçimi isə sorğu

olunan cümlələrin seçilməsidir. Biz alqoritm bir az dəyişəcəyik. Biz informativ olan sözləri

ya loqarifma ehtimal nisbəti ya da sorğuda görünmələrinə görə seçəcəyik. Biz sənəddəki hər

bir sözə çəki əmsalı verəcəyik. Biz sözə loqarifma ehtimal nisbətinin limitini,

yəni 10 limitini keçəcəksə, 1 çəki əmsalı verəcəyik. Biz həmçinin sözə əgər söz bizim

göndərdiyimiz sorğuda da varsa 1 çəki əmsalı verəcəyik. Digər halda söz 0 çəki əmsalı alacaq.

Bu çəki əmsalları çox sadədir, sadəcə 1 və 0. Siz daha mürəkkəb çəki əmsallarını da öyrənə

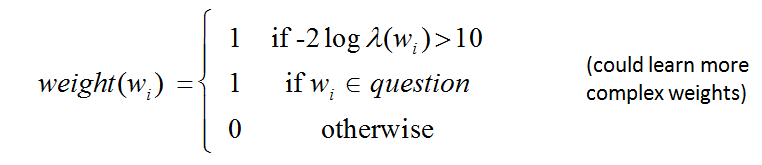
bilərsiz. Və daha detallı çəki əmsallarını öyrənməyin güclü yolları təyin etmək üçün elmi-

araşdırma gedir. Ancaq, 1 və 0 normal işləyir. İndi, biz cümlənin çəki əmsalını hesablayacağıq,

bəlkə də pəncərini öyrənəcəyik harda ki, bizim fakitiki cümlələrimiz yoxdur. Biz cümləni çəki

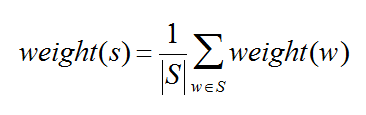
əmsalını sözlərin çəki əmsallarının cəmini hesablayıb ədədi orta tapmaqla tapacağıq.

* İnformativ sözləri seçmək üçün
  + Loqarifma ehtimal nisbəti (LEN) ilə
  + Sorğuda görünməsinə görə



* Cümlənin (və ya pəncərənin) çəki əmsalını sözlərin çəki əmsalları

vasitəsilə tap:



Bizim mövzü seçmə alqortmimiz müəllimsiz öyrətməyə aiddir. Bizim qısa

məzmunlardan ibarət nişanlanmış öyrədici bazamız yoxdur ki, biz onun

vasitəsilə çəki əmsallarını öyrənək. Bu halda bizim alternativ yanaşmamız

var: müəllimli mövzu şeçilməsi. Əgər bizim yaxşı qısa məzmunlu

sənədlər və məzmunlarda hər bir cümlə üçün sıralanma olan

nişanlanmış bazamız varsa və əgər biz hansı cümlənin sənəddən gəldiyini

biliriksə o zaman bizim uyğun cümlələrimiz var. Və biz istənilən tip xüsusiyyəti

çıxarda bilərik. Məsələn cümlənin sənəddəki yerini tapa bilərik. Birinci cümlələr

uzunluqdan asılı olmayaraq çox zaman yaxşı qısa məzmun cümlələri olur.

Biz əvvəl sahib olduğumuz bütün xüsusiyyətləri əldə edə bilərik, misal üçün

sözün informativliyi və s. Nitq informasiyasına əsaslanan digər növ xüsusiyyətlərimiz

də ola bilər. Biz hər bir cümləni hər hansı xüsusiyyətlər vektoru ilə

əlaqələndirə bilərik və biz ikili sinifləndiricini öyrədə bilərik. Mən bu cümləni qısa

məzmunda yerləşdirəcəyəmmi? Hə və ya Yox. Və o çəki əmsallarını bütün bu

xüsusiyyətlər üçün və ya yaradılmış yeni xüsusiyyətlər üçün öyrənə bilər. Alqoritm

üçün bu yaxşı səslənə bilər, ancaq praktikada bu tipli nişanlanmış bazanın tapmaq

çox çətindir. İnsanlar xülasə yazanda həmişə sənədin özündə olan

eyni sözlərdən, ifadələrdən və tam cümlələrdən istifadə etmir. Beləliklə, sənəddən

çıxarılmış mükəmməl nişanlanmış xülasələri tapmaq çətindir. Eyniləşdirmə etmək

də çətindir, çünki, onlar tam cümlələri deyil, ya sözləri ya ifadələri ya da parçalari

götürürlər. Sözləri sənəddən götürmələrinə baxmayaraq həmin sözlərin sənədin hansı

hissəsindən gəldiyini tapmaq çətindir. Və sürpriz haldır ki, performansı müəllimsiz

öyrətmə alqoritmindən o qədər də yaxşı deyil. Praktikada ancaq loqarifma ehtimal

nisbəti və sözün və ya cümlənin hansı dərəcədə səciyyəvi və ya infromativ olduğunu

təyin edən digər ölçmələrdən istifadə edən müəllimsiz məzmun seçmə daha çox tətbiq

edilir. Biz tək sənəddən necə qısa məzmun çıxartmanı gördük. Götürdüyümüz

əsas alqoritm sadə statistik yolla informativ sözlərə baxmaqla bizə lazım olan cümləni

tapmağa kömək etdi. Və bu cümlələrin tapılmasında loqarifma ehitmal nisbətinin önəmli

bir üsul olduğundan danışdıq.