SVEUČILIŠTE U SPLITU

PRIRODOSLOVNO-MATEMATIČKI FAKULTET

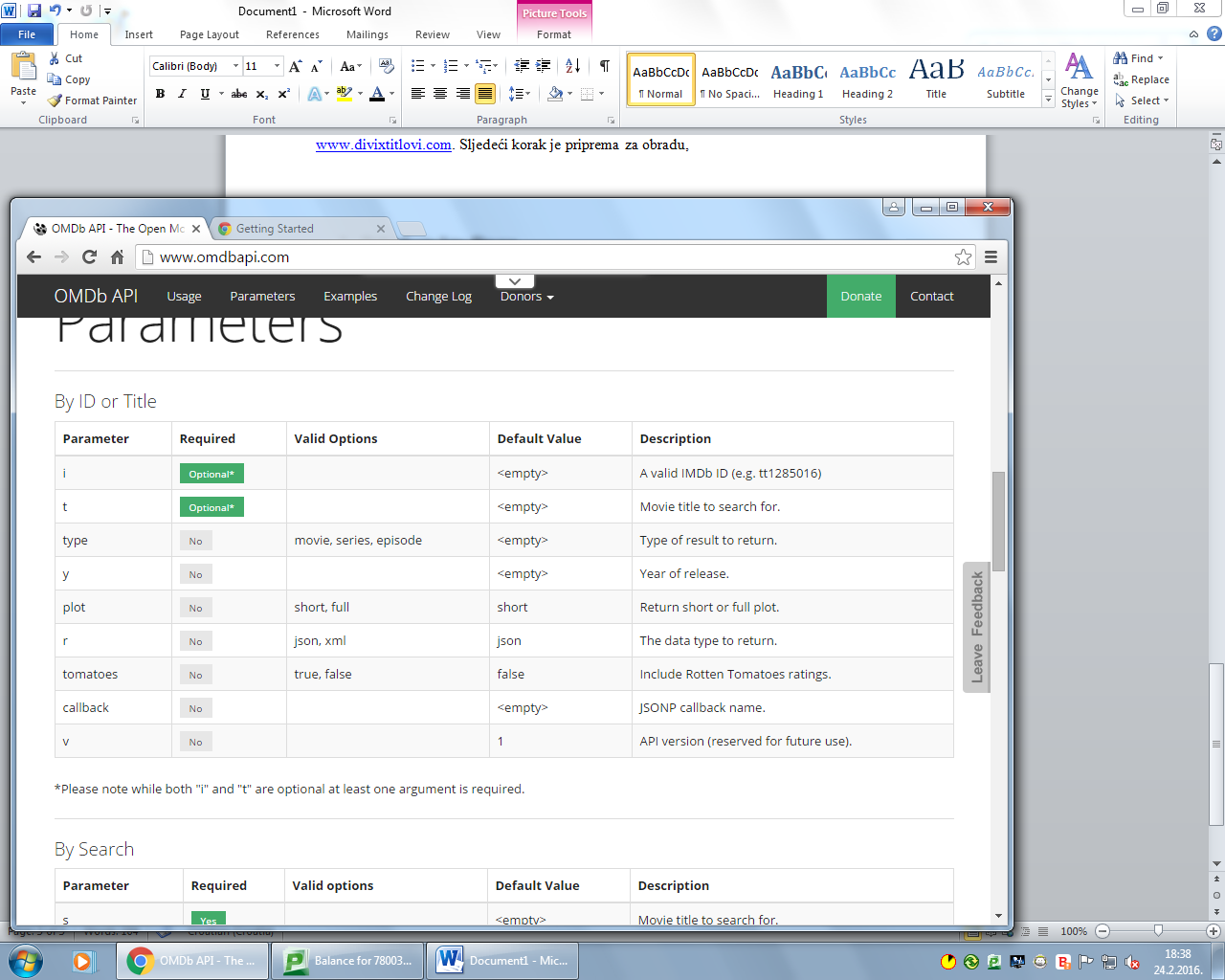
Sadržaj :

1. Uvod
2. Sastavljanje liste filmova
   1. Preuzimanje titlova
3. Obrada
4. **Uvod**

Cilj ovog projekta je napraviti klasifikaciju titlova po žanru filma. Svaki film može imati jedan ili više žanrova Prvi korak u izradi je prikupljanje liste filmova. Nakon toga spajanje na API i preuzimanje titlova sa stranice [www.divixtitlovi.com](http://www.divixtitlovi.com). Sljedeći korak je priprema za obradu. Kao mogući žanrovi filmova su ponuđeni: ( action, horror, sci-fi, war, sport, romance, western, crime ). Žanrove comedy, drama, history, thriler namjerno nismo koristili jer gotovo uvijek dolaze u kombinaciji s nekim drugim žanrom.

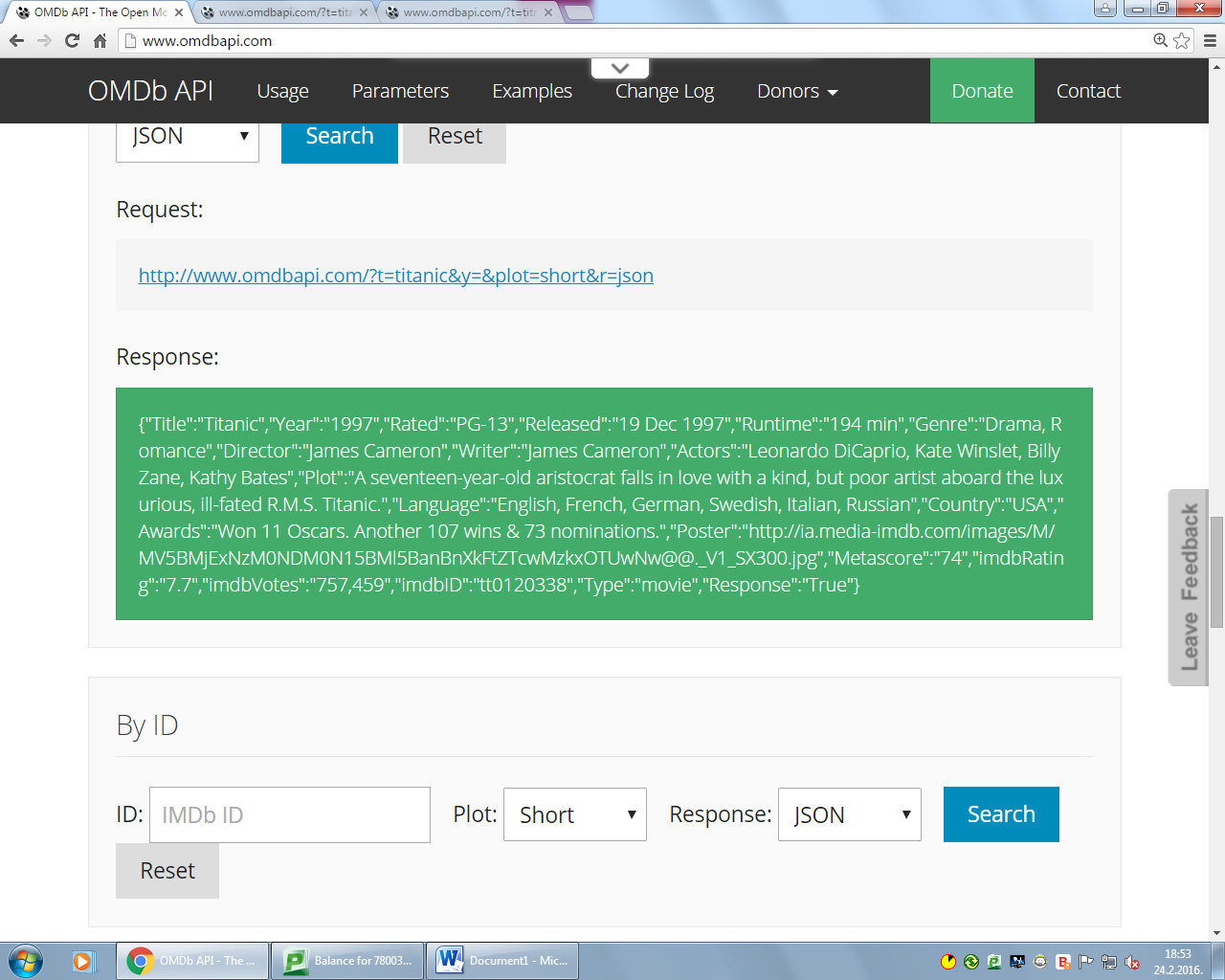
1. Sastavljanje liste filmova

Za dobavljanje liste filmova se koristi API koji omogućuje brzo i efikasno preuzimanje većeg broja titlova sa stranice „The Open Movie Database“ (OMDb). Stranica omogućuje jednostavne upite koja daje informacije o filmovima prema stanju na najpoznatijoj bazi podataka filmova IMDB.



Slika 1.

Na slici 1. su prikazana polja po kojima možemo pretraživati filmove. Najbitnije polje za dobavljanje liste filmova je ID, koji predstavlja identifikacijski broj svakog pojedinog filma. Informacije o filmu se dobivaju tako da korisnik šalje serveru upit na koji server odgovara porukom o informacijama koja je u obliku:



*Slika 2. Upit i odgovor*

Kako bi prikupili što više filmova poslužili smo se malom mudrolijom, shvatili smo da se ID filmova kreće u rasponu od 100 000 do 999 999. S obzirom na to pretraga po slučajnom broju u tom rasponu u više od 50 % slučajeva daje pozitivan rezultat tj. vraća informacije o nekom filmu u bazi. Baza podataka sadrži mnogo neiskoristivih unosa, ali unatoč tome slanje uzastopnih upita koristeći for petlju se može doći do valjanih informacija o filmovima.

Paralelnim pokretanjem skripte „**spajanje\_na\_api\_zafilmove.py**“ smo došli do više disjunktnih listi informacija o filmovima u obliku .txt zapisa na način **„Naslov\*ID\*Žanr“.**

* 1. **Preuzimanje titlova**

Za preuzimanje titlova se koristi skripta „**preuzimanje\_titlova.py**“. Skripta funkcionira na način da za svaku liniju iz liste filmova filtrira valjanje linije, nakon toga provjerava je li postoje titlovi za film. Ukoliko postoje titlovi preuzima ih i sprema sve potrebne informacije u datoteku **filmovi\_info.txt oblika „Naslov filma\*Putanja titlova\*Žanr“**.

1. **Obradu**

U ovoj fazi potrebno je odrediti način na koji će se odrediti žanr filma. Koristi se algoritam napravljen od 2 djela rangiranja koji za određene titlove daje ocijenu za svaki mogući žanr. Nakon što algoritmi završe s obradom bira se žanr s najvećom ocijenom. Za funkcioniranje oba algoritma je potreban **skup specifičnih riječi** za svaki žanr. Skupovi specifičnih riječi se nalaze u direktoriju **zanrovi**.

Odabiranje specifičnih rječi za pojedini žanr je vrlo složen proces jer se svaka riječ može upotrebljavati u različitim kontekstima. Za odabir rječi smo se služili skriptom pred\_obrada u kojoj bi vidjeli skup riječi koje se najviše ponavljaju u određenim žanrovima te bi među njima odabrali one koje najbolje opisuju određeni žanr. Za drugi algoritam se radi isti proces samo se barata s dvije uzastopne riječi, te se tako dobiva malo više značenja. Jednom kad se odredi skup riječi on ostaje konstantan dio programa. Odabiru se isključivo kombinacije dvije riječi gdje je bar jedna sadržana u skupu specifičnih rječi. Svim elementima koji rade sa kombinacijma dvije rječi je na kraj dodan nastavak „**\_X2**“ radi raspznavanja. Na taj način se sve više približavamo kontekstu specefične riječi te je algoritam moguće dodatno nadograditi da gleda i veće strukture koje bi našao preko manjih.

U konačnoj obradi se računa koliko se puta svaka riječ iz specifičnog skupa pojavljuje u titlovima. Svako pojavljivanje rječi se bilježi, pa na kraju vidimo točan broj pojavljivanja svake riječi u titlovima. Slijedi rankiranje svakog žanra gdje se rank dobiva na sljedeći način:

Rank(žanr) = suma spec. riječi(žanr) + broj različitih riječi koji se pojavljuje + (suma spec. riječi\_X2(žanr) \* 5)

Nakon toga slijedi odabir žanra s najvećim rankom.