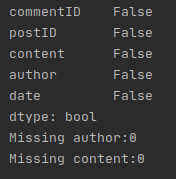
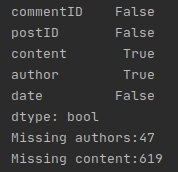
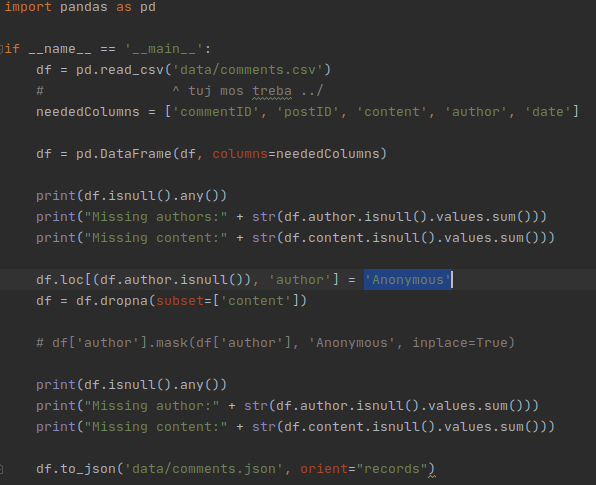
**Предпроцесирање**

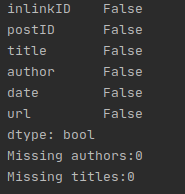
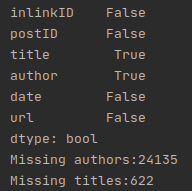
Во оваа фаза податоците ги спремивме за подетална обработка така што таму каде што недостасувашеа вредности во некои колони или ги потполнивме со некаква default-на вредност или пак го избришавме целиот ред доколку податоците немаат смисла без тој податок.

За табелата коментари, авторот на коментарот недостасуваше на 47 места, додека самата содржина на коментарот недостасуваше 619 пати. Вредностите каде што недостасуваше авторот ги потполнивме со 'Anonymous', но бидејќи нема смисла да чуваме податоци за коментар кај кој нема неговата содржина затоа го избришавме целиот ред за тој коментар.



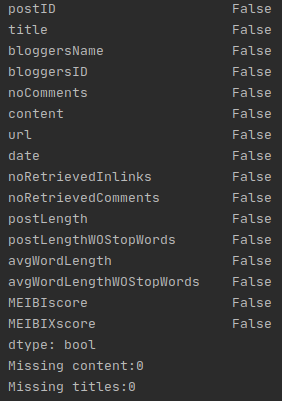
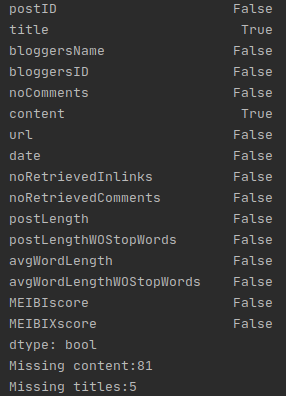


За табелата inlinks, авторот недостасуваше 24135 пати и 622 пати недостасуваше насловот. Бидејќи авторот и насловот не се битни за еден inlink, вредностите каде што недостасувашеа за авторот ги потполнивме со 'Anonymous' и за насловот со 'No title'.



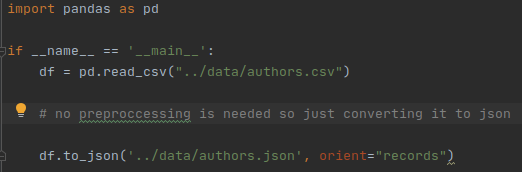


За табелата posts, насловот на post-от недостасуваше на 5 места и самата содржина за насловот недостасуваше 81 пат. Вредностите каде што недостасуваше насловот ги потполнивме со 'No title' бидејќи може да има пост кој нема наслов но има содржина. Таму каде што немаше содржина запишавме 'No content available' бидејќи можеби насловот е самоопишувачки.

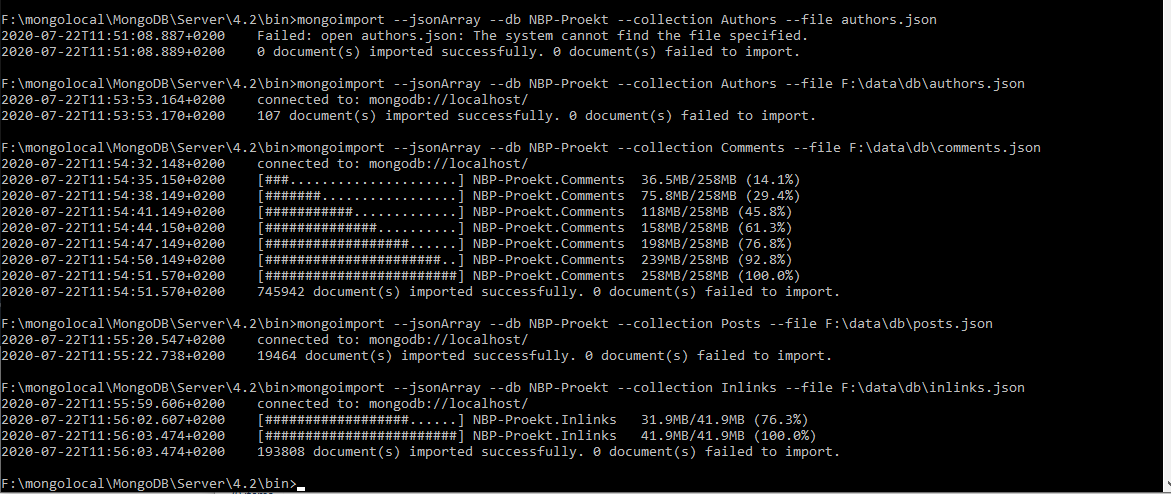
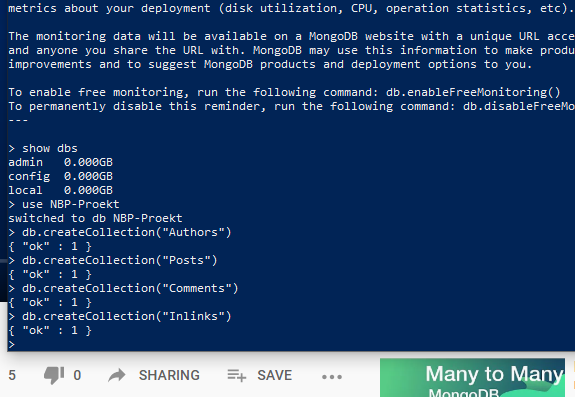




Табелата за авторите нема вредности кои недостасуваат па затоа таа табела остана иста.

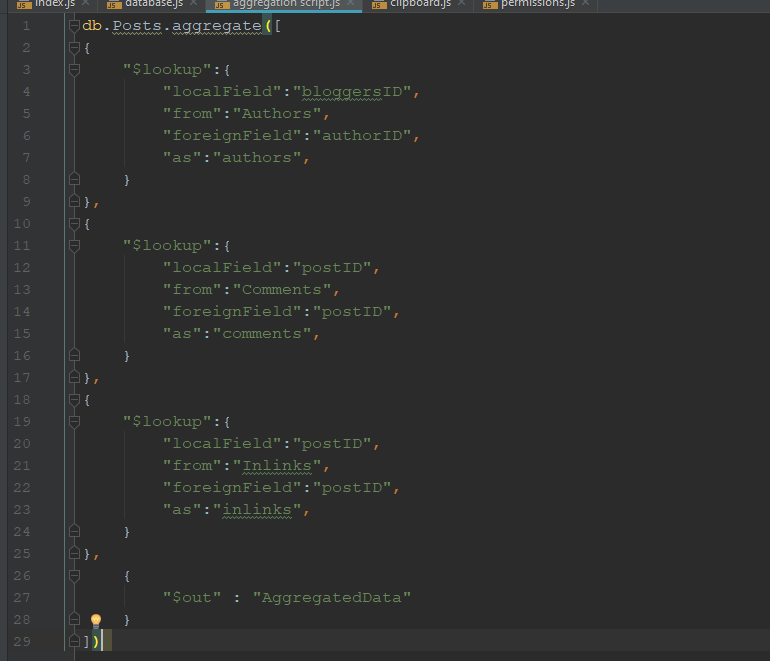


**Внесување на податоците во базата**



**Креирање на првата структура на податочното множество**

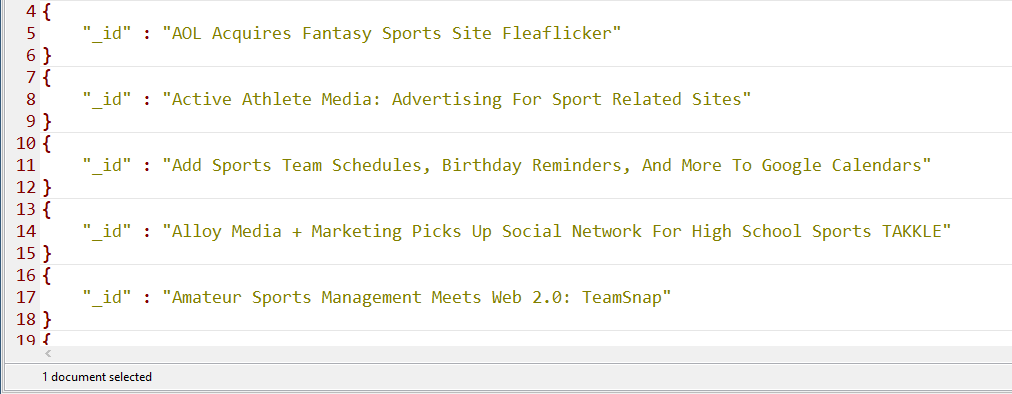
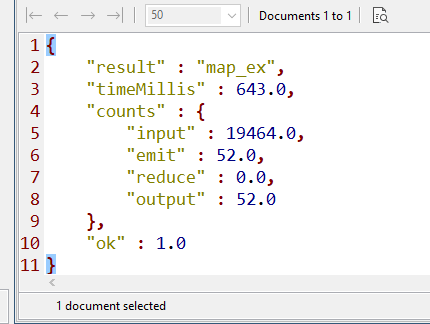
Како прв начин за организација на податотоците ни беше на врвот на хиерархијата да се наоѓаат податоците за Posts, а потоа во нив да се вгнездени сите податоци за коментарите, авторите и инлинковите.Ова го постигнавме со следниов код за агрегација на податоците.



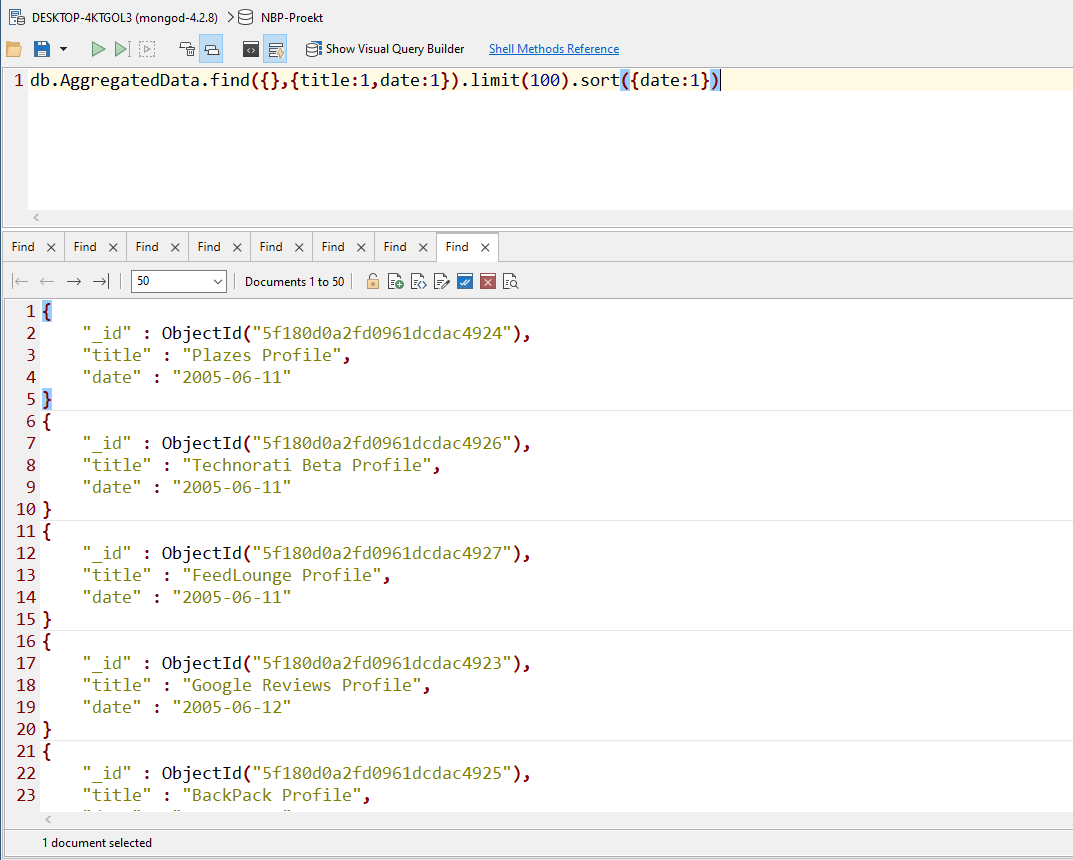
Времето потребно за да се изврши оваа агрегација беше околу 2-3 часа. Немаме нај прецизна информација бидејќи заборавивме да измериме.

**Поставување на прашања до базата со map-reduce и без map-reduce.**

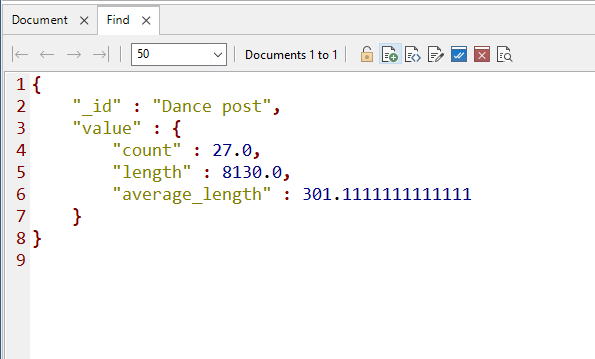
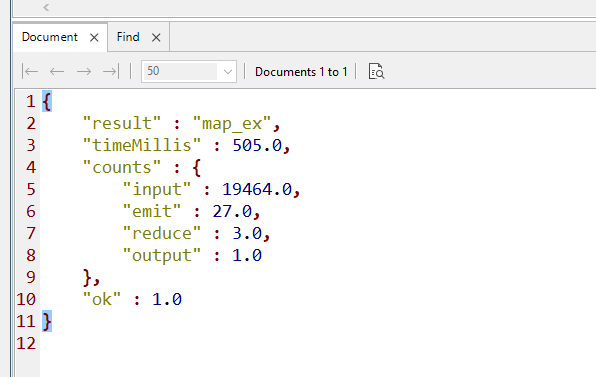
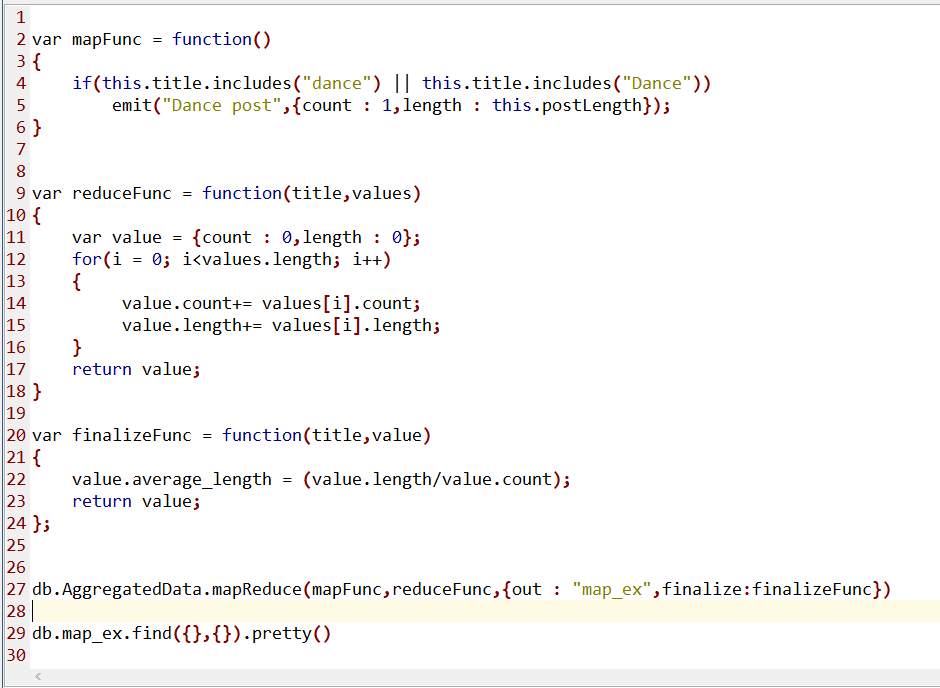
1.Најди ги сите наслови на објави кој се поврзани со спорт.



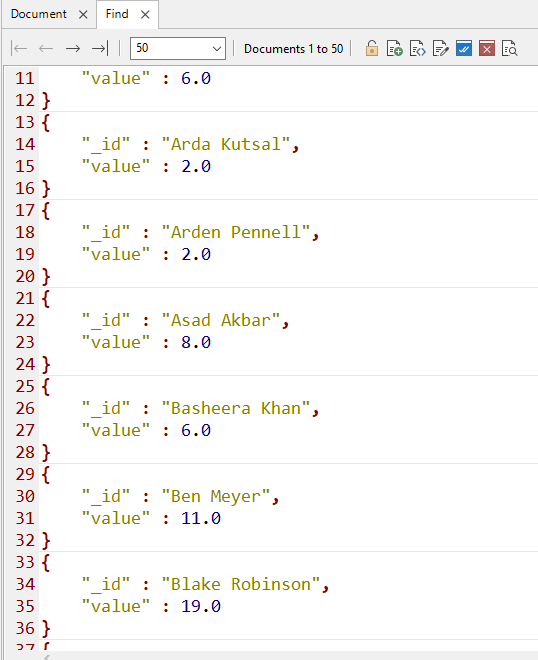
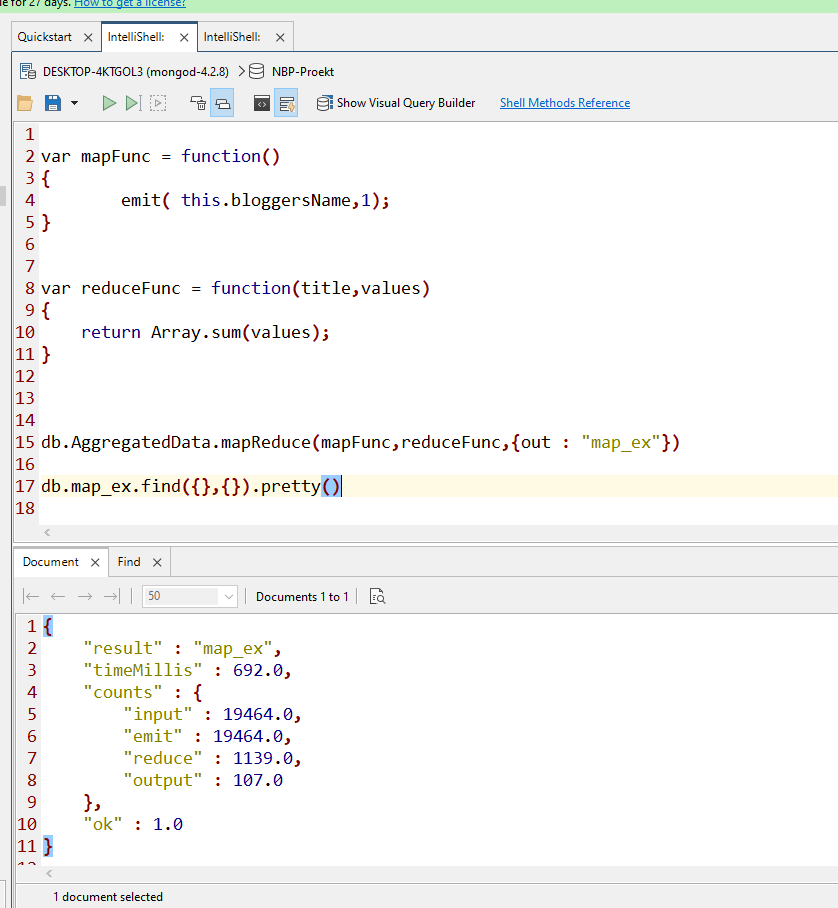
2.Дај го насловот и датумот на најстарите 100 објави.



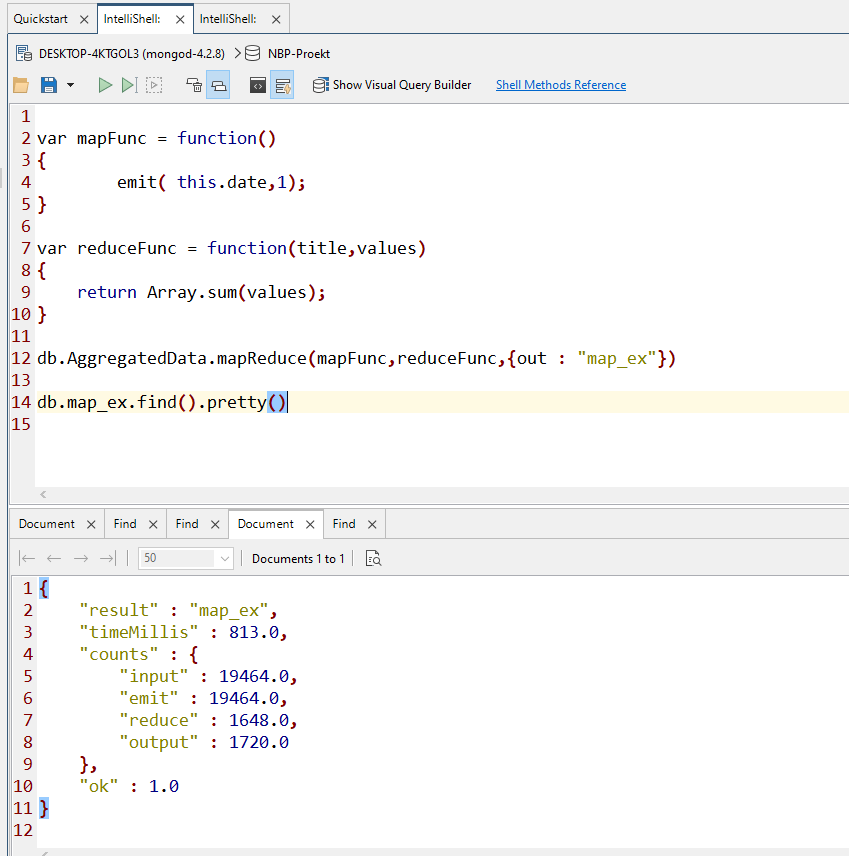
3.Дај ја просечната должина на постовите кои се во врска со танцување(dance).

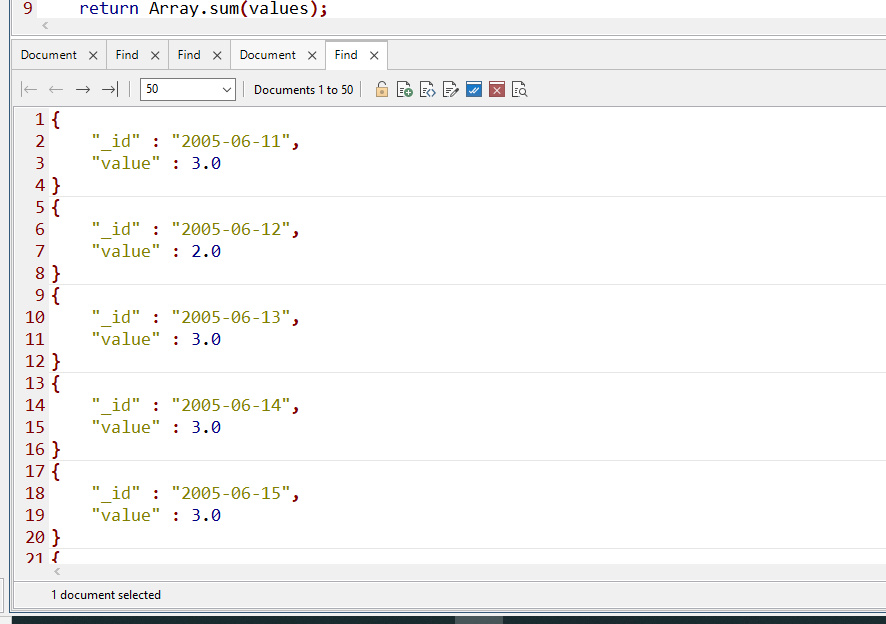


4.Најди кој автор колку постови има објавено.

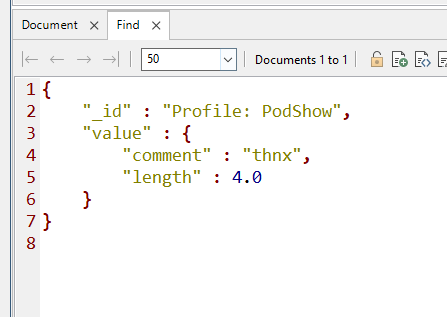


5.На секој датум колку постови се објавени.





6.Најди го најкраткиот коментар во сите објави заедно.



7.Најди го авторот со највеке коментари на објавите и колку коментари во просек има.

