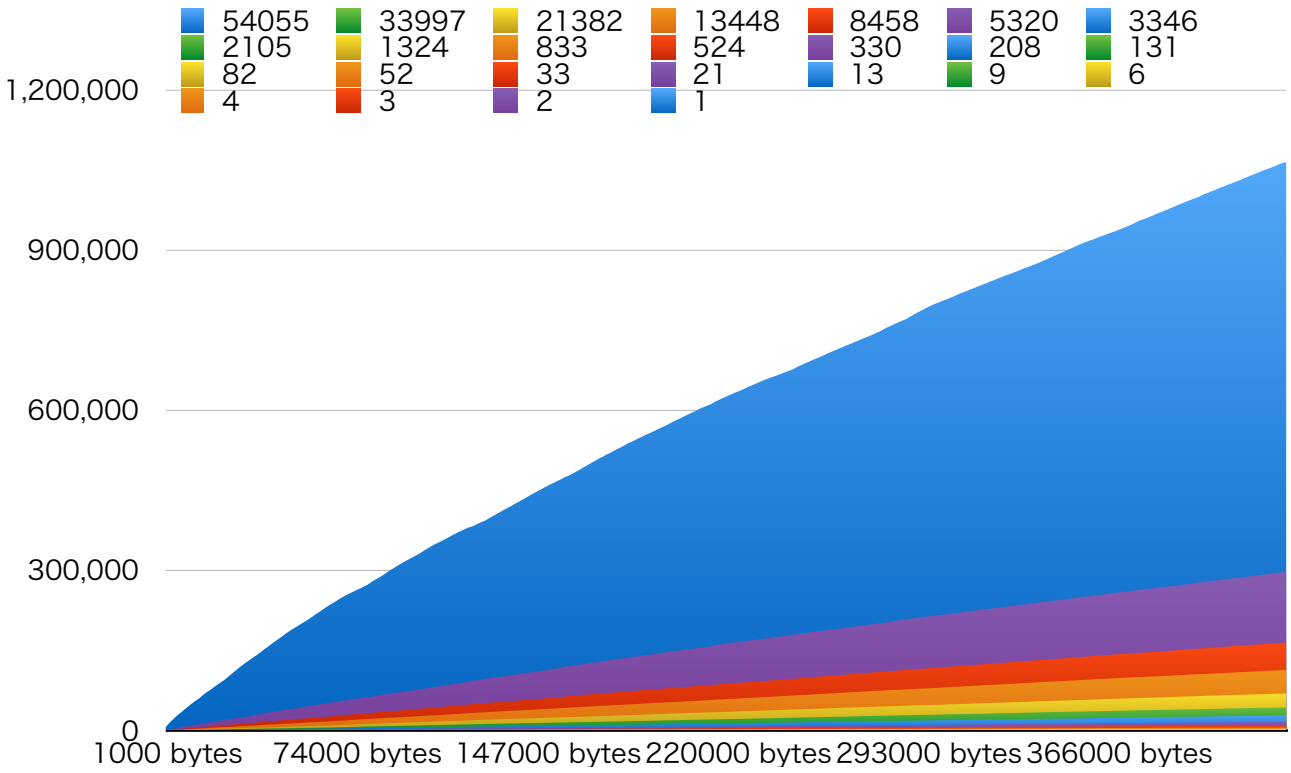
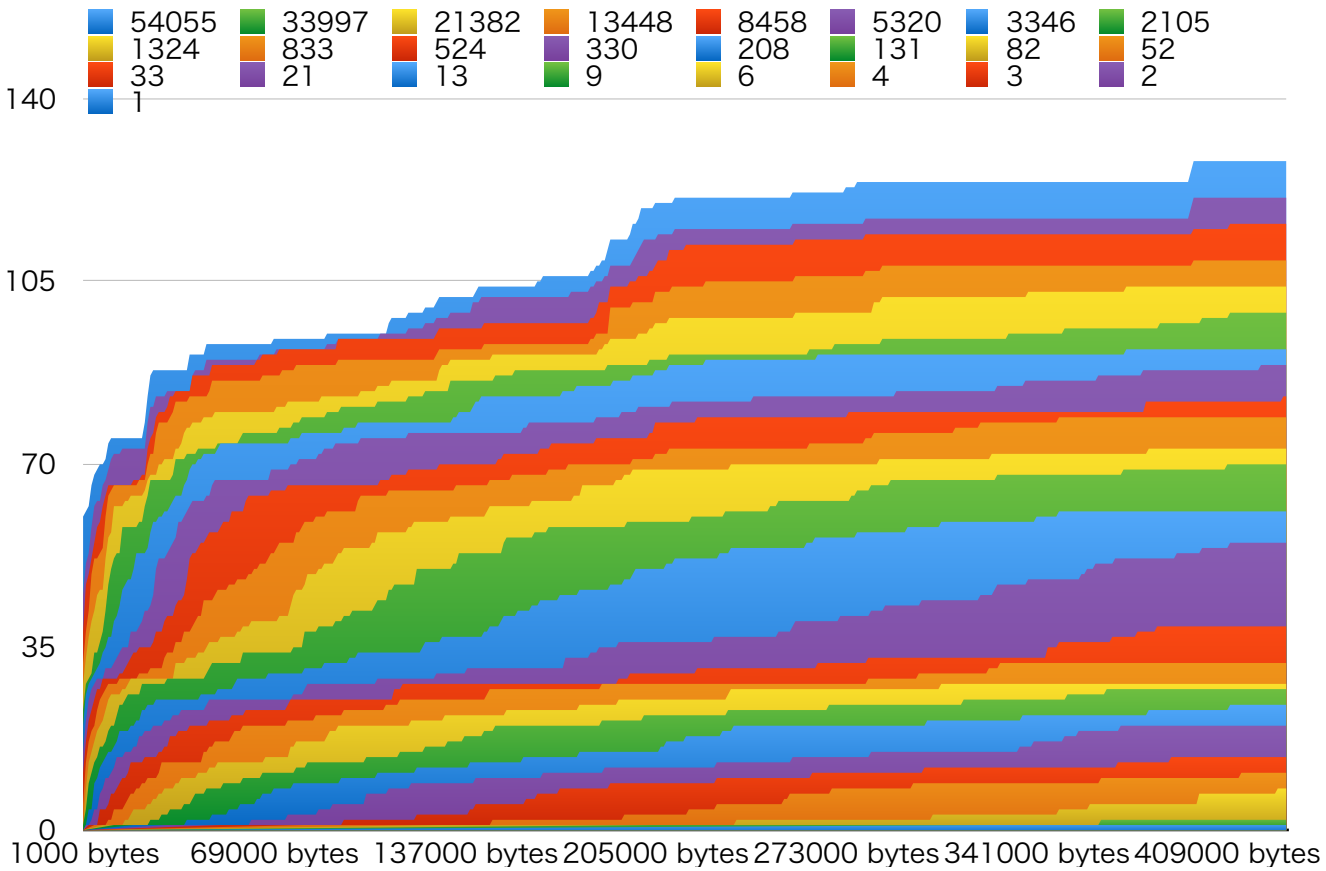
n-grammien määrän kasvu oppimisdatan kasvun suhteen (430 KB, englanninkielisellä datalla)

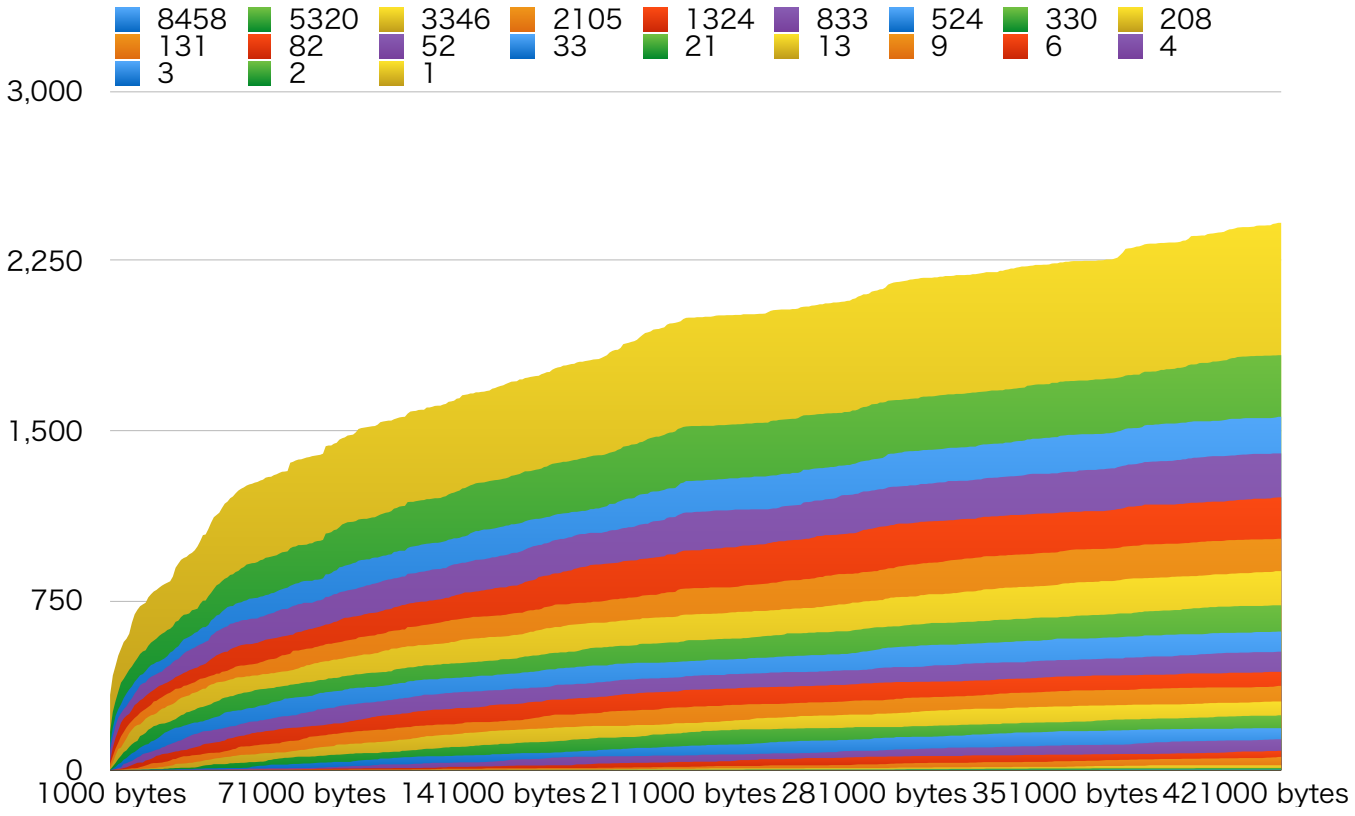
n-grammit (kaikki yhteensä)



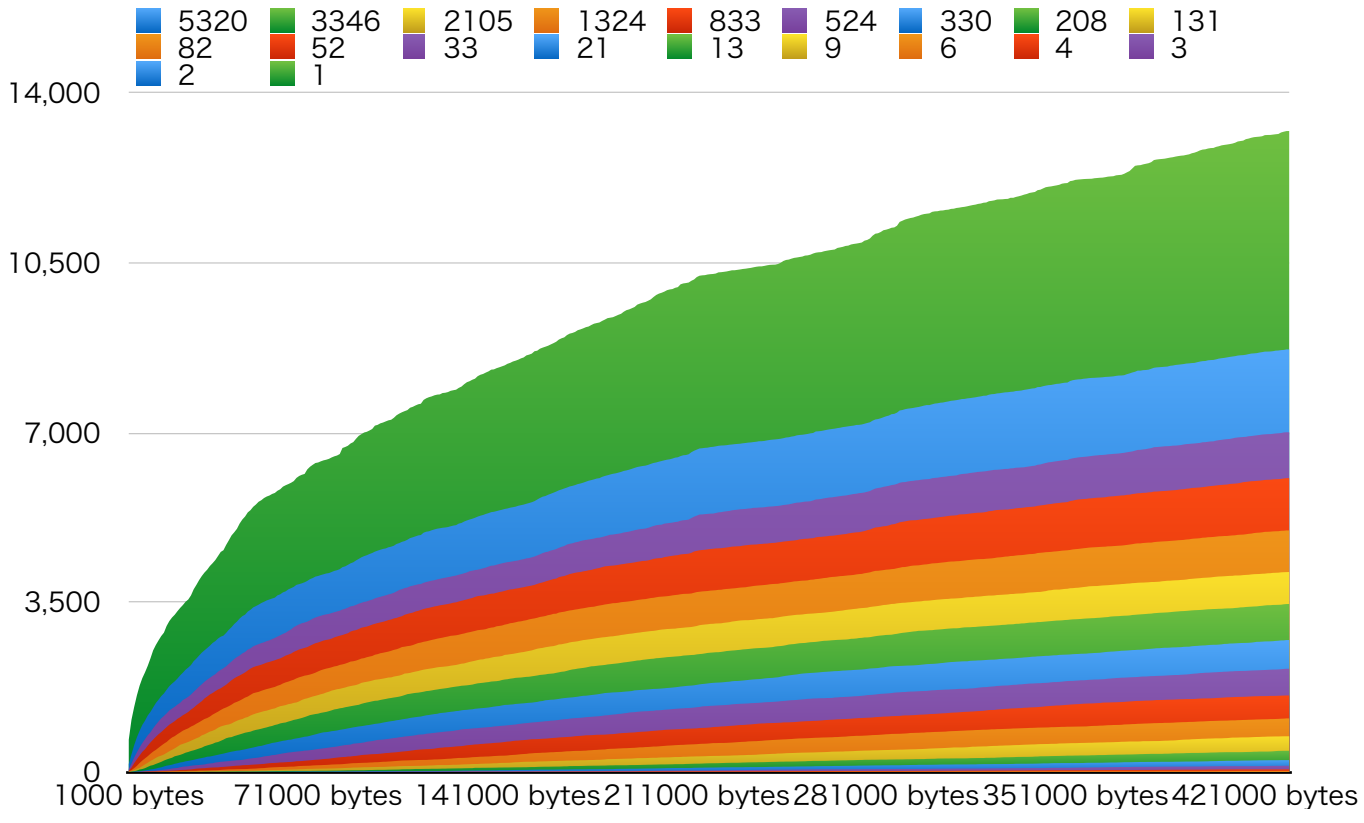
1-grammit



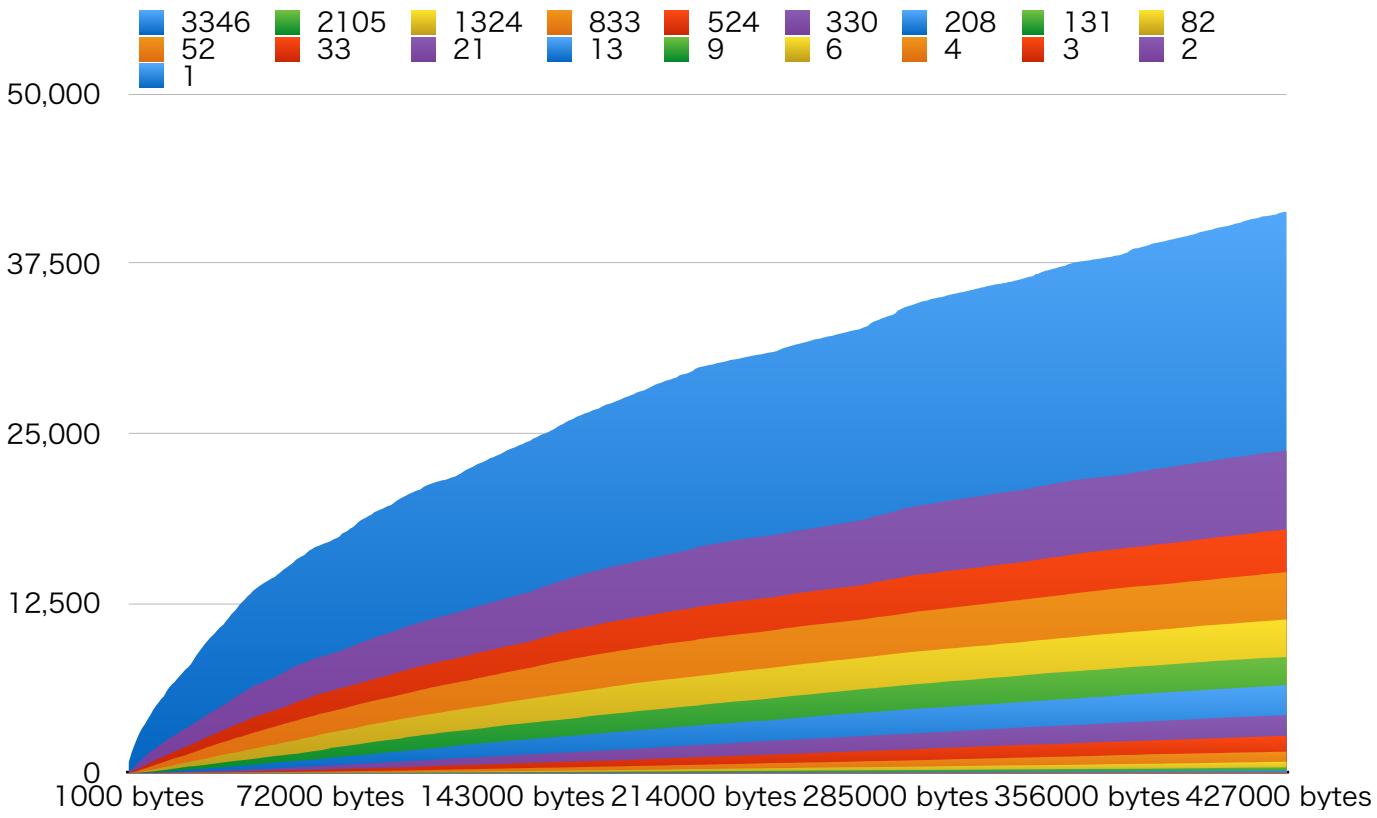
2-grammit



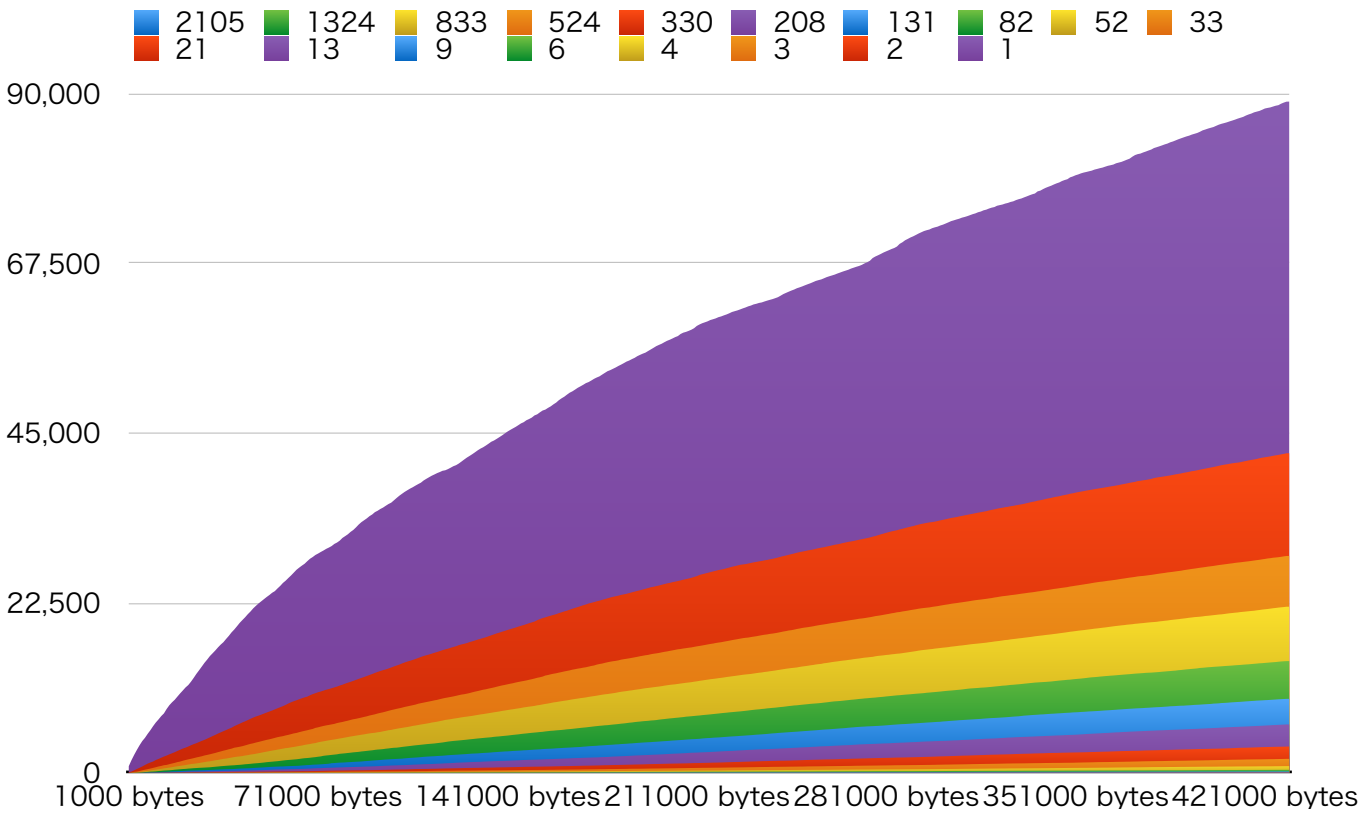
3-grammit



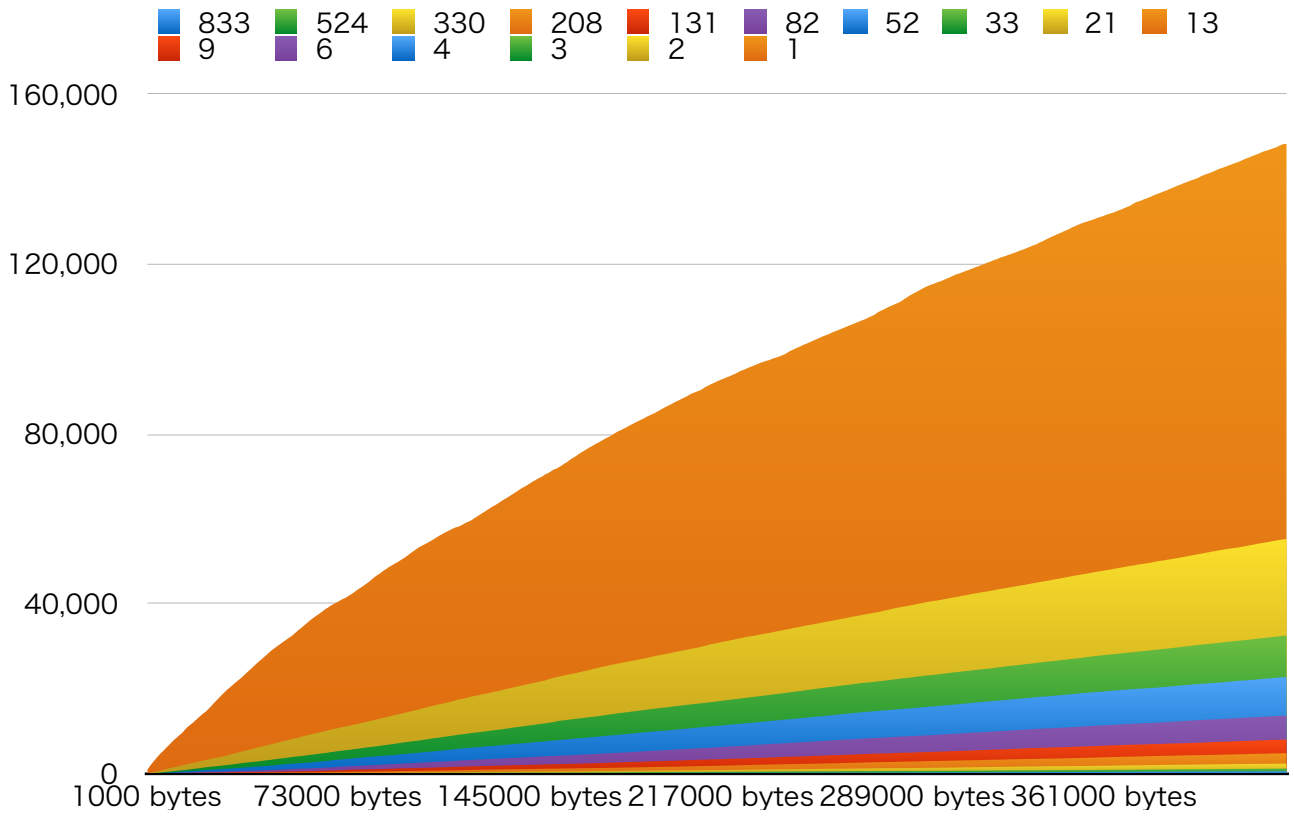
4-grammit



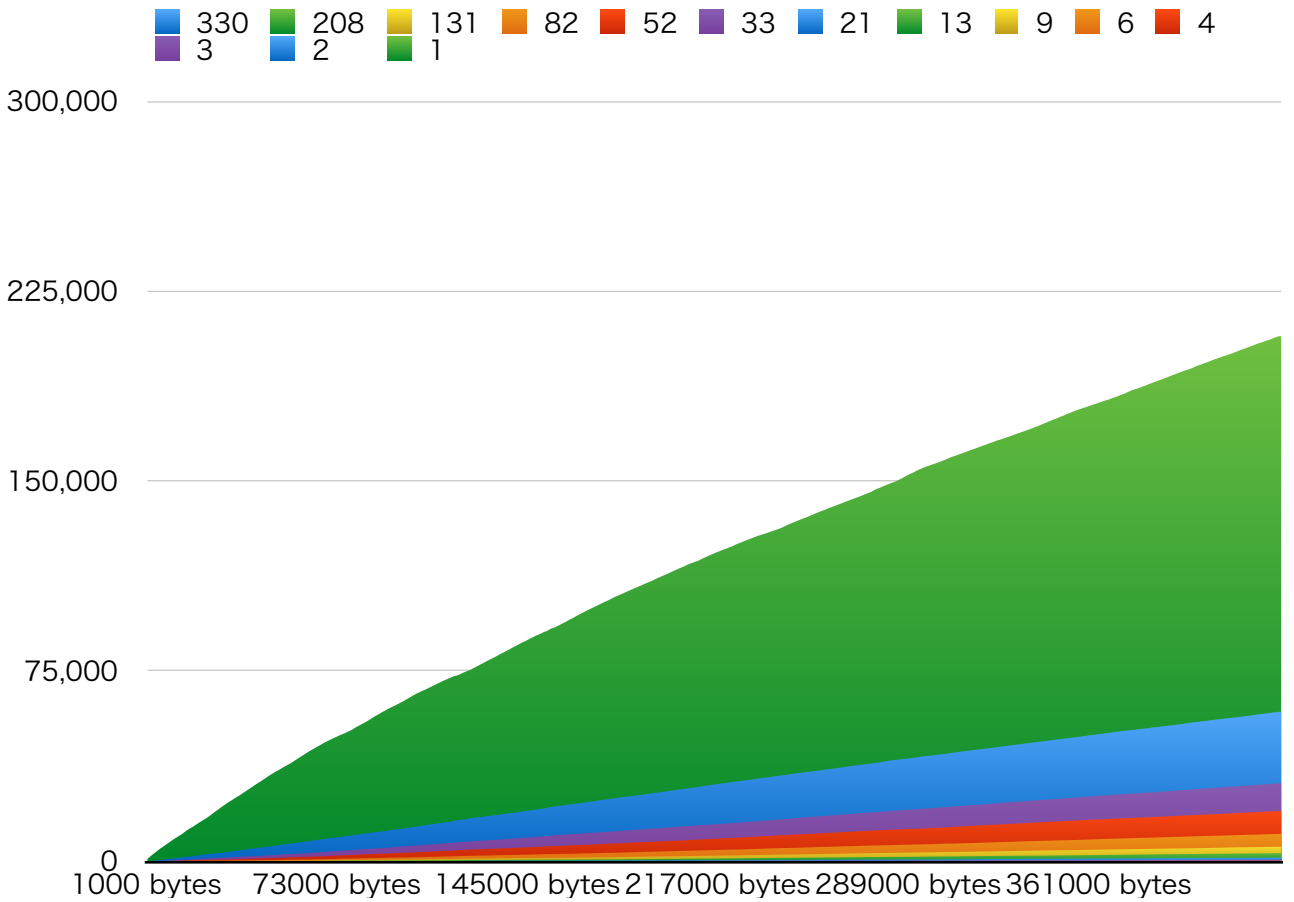
5-grammit



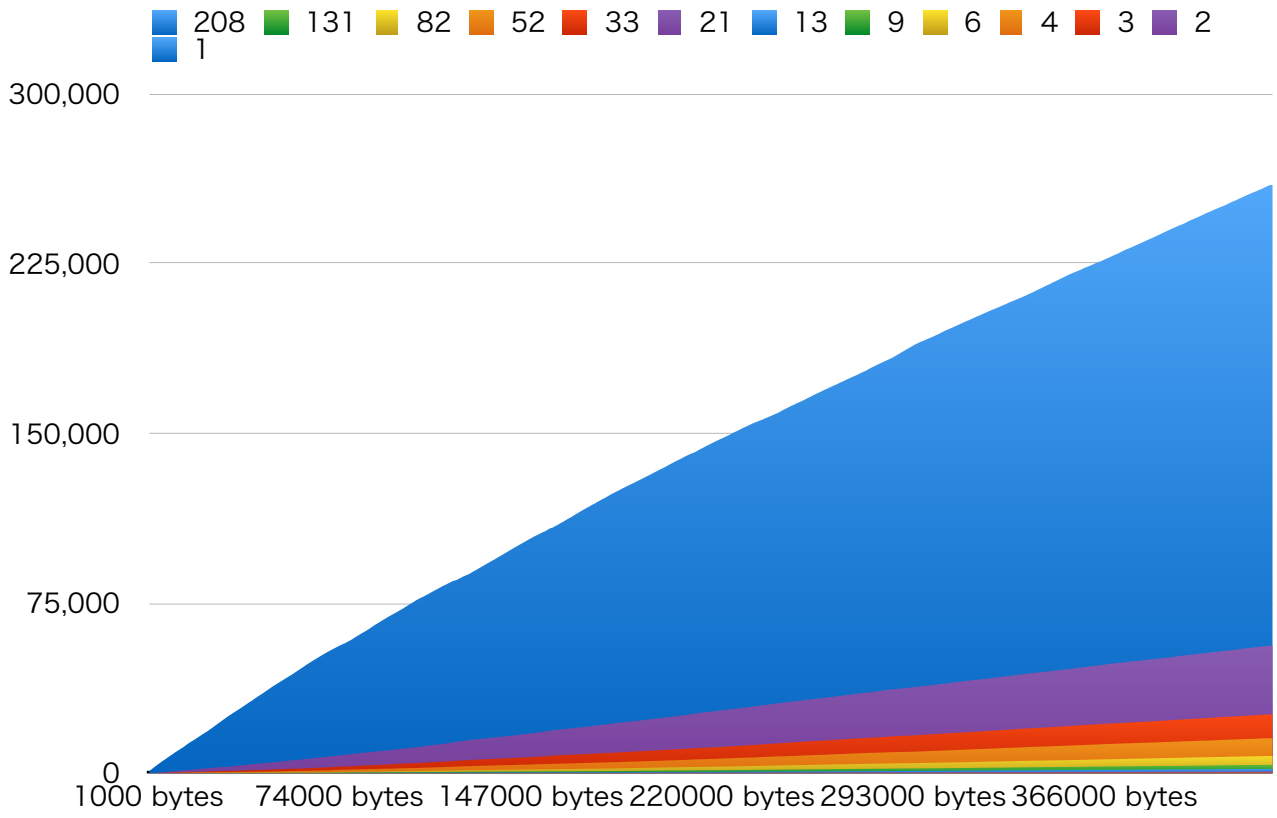
6-grammit



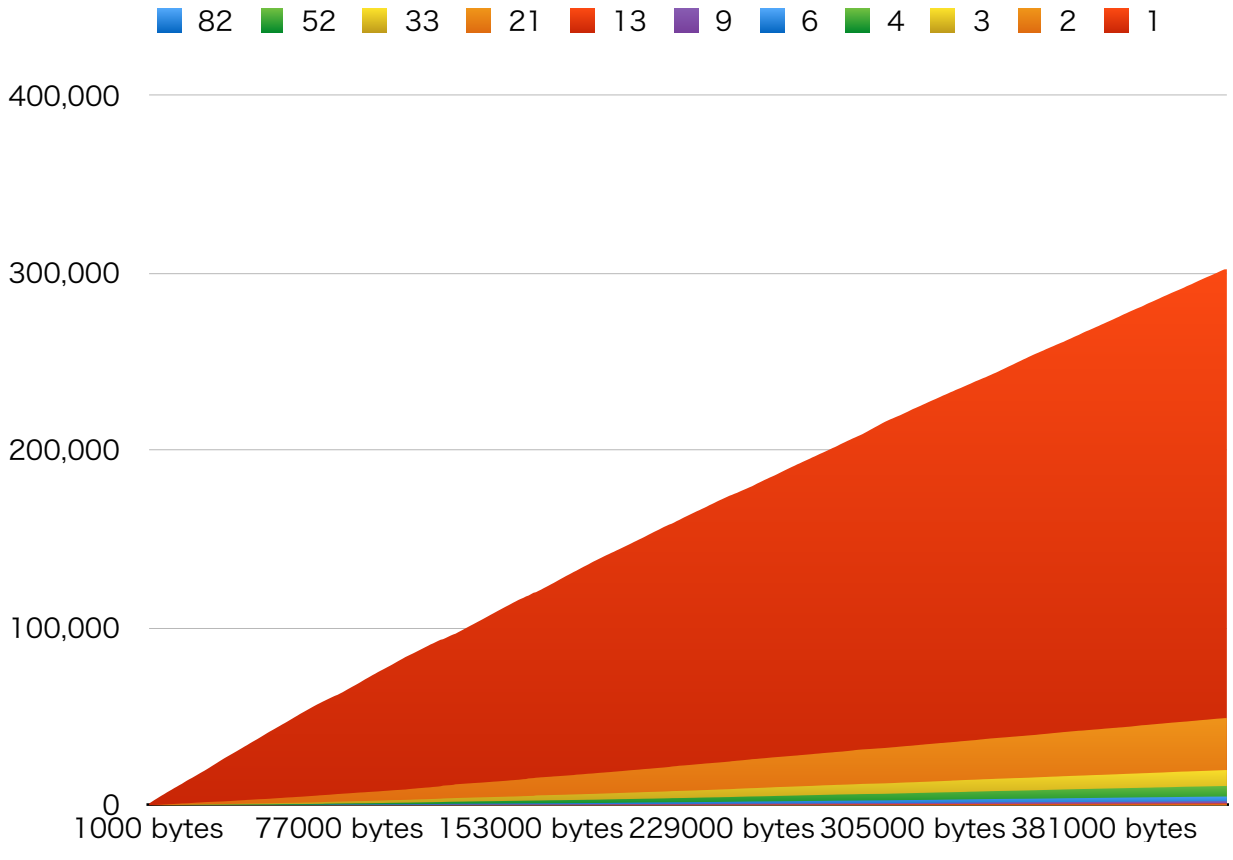
7-grammit



8-grammit



9-grammit



Tästä datasta näemme sen (luonnollisen) päätelmän, että mitä pidempiä n-grammit ovat, sitä hitaampaa on niiden määrän kasvun taittuminen. 1-grammit, eli yhtä tavua per merkki käyttävän englannin kielen tapauksessa yhden kirjaimen grammien määrä kasvaa nopeasti niin korkeaksi kuin on mahdollista, kun taas pitemmillä n-grammeilla kasvu jatkuu paljon pidemmälle. (Eri värit viittaavat siihen, kuinka monta kertaa aineistossa kys. n-grammi mainitaan. Luonnollisesti vain yhden kerran mainittuja on varsinkin pitkillä n-grammeilla eniten.)

Kuinka suuri määrä oppimisdataa tarvitaan, että saadaan tarkka kuva kielen n-grammiprofiilista? Eräs keino arvioida asiaa lienee juurikin n-grammien kasvun tarkkailu oppimisdatan määrän suhteen: jossain vaiheessa kokoelmamme n-grammeja alkaa ”saturoitua”. Saturoituminen yllä olevassa aineistossa on selvää vielä 4- ja 5-grammeilla.

Jo pelkillä kirjainprofiileilla eli 1-grammeilla pystytään päättelemään kieli jossain määrin, ja jokainen taso tähän päälle parantaa mallin tarkkuutta. Englannissa on keskimäärin 4,5 kirjainta per sana (lähde: John R. Pierce - An Introduction to Information Theory: Symbols, Signals and Noise), joten 5-grammit riittävät tunnistamaan jo keskimääräisiä englanninkielisiä sanoja. Kunhan varsinainen kielenarvausalgoritmi saadaan implementoitua, voidaan verrata varsinaisia tuloksia erilaisella oppimisdatalla.