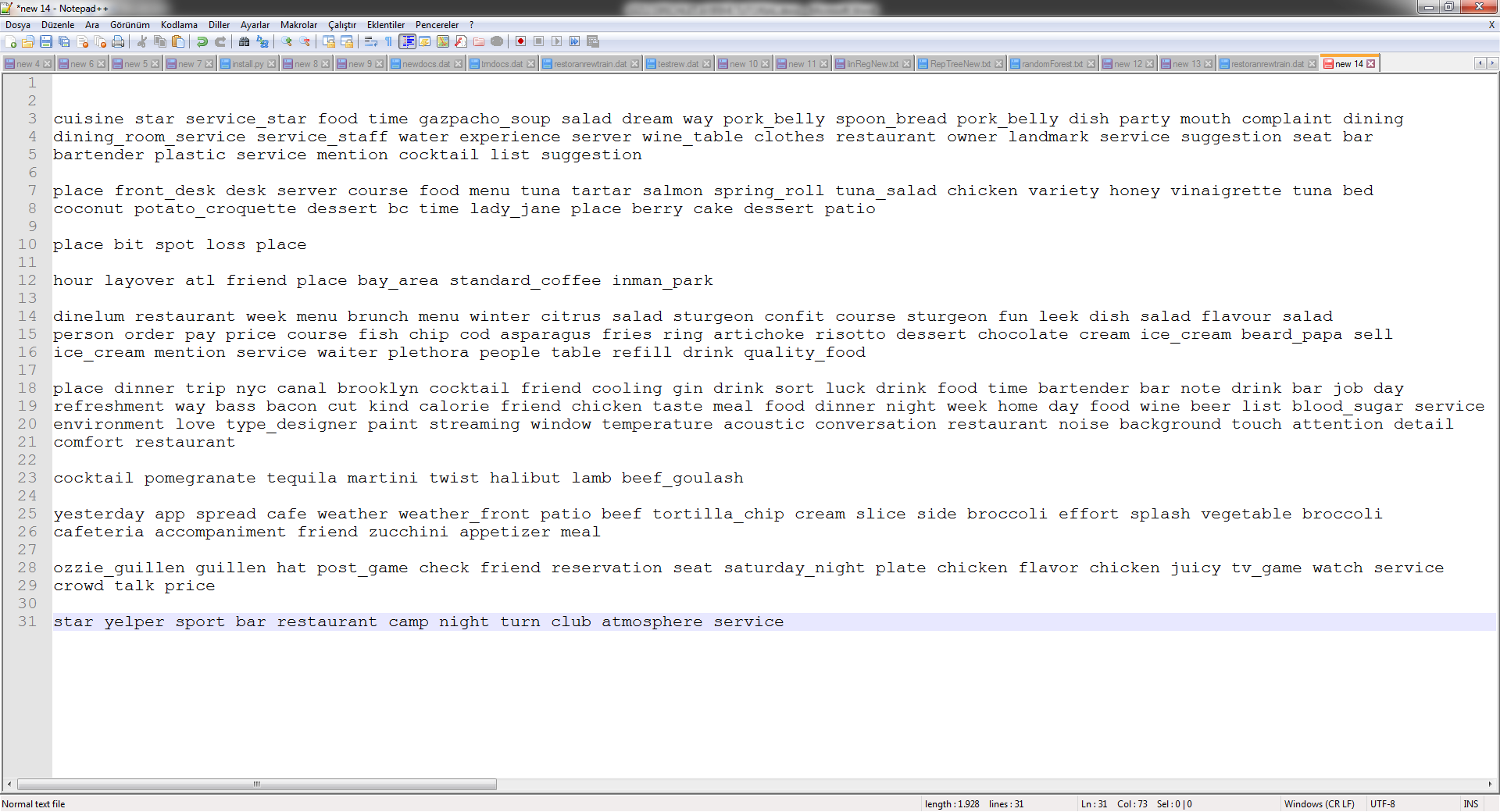
# GİZLİ DİRİCHLET AYIRIMI

Bu dokümanda Gizli Dirichlet Ayırımı örnek bir yorum kümesi üzerinden adım adım anlatılmaktadır.

Gizli Dirichlet Ayırımının uygulanacağı koleksiyon 10 adet yorumdan oluşmaktadır. Koleksiyon Şekil 1’de verilmiştir.



**Şekil 1. Örnek Koleksiyon**

GDA’daki temel varsayım şu şekilde özetlenebilir: 1) bir doküman birden fazla konu içermektedir, 2) konular kelimeler üzerinden dağılıma sahiptir, 3) her doküman için konu dağılımları belirlenir, 4) dokümandaki her kelime için konu dağılımları üzerinden bir konu seçilir. GDA’nın eldeki dokümanlara uygulanması aşağıda anlatılmıştır.

1. GDA’nın eldeki koleksiyona uygulanabilmesi için ilk adımda model parametrelerinin belirlenmesi gerekmektedir. Bu parametrelerden ilki gizli konu sayısı ’dır. Bu koleksiyon için  olarak belirlenmiştir. Dirichlet parametreleri  ve ‘nın başlangıçta tanımlanması gerekmektedir.  konuların dokümanlardaki dağılımı ile ilgili parametre olup, değeri olarak atanmıştır.  ise kelimelerin konulardaki dağılımı ile ilişkili parametre olup değeri 0.01 olarak atanmıştır. İterasyon sayısı da 10 olarak belirlenmiştir.
2. İkinci adımda koleksiyonda yer alan kelimelerden bir sözlük oluşturulur. Sözlükte toplamda 187 farklı kelime/kelime öbeği bulunmaktadır.
3. GDA ilk olarak her doküman için dokümandaki kelimelere rastgele konu ataması yaparak başlar. Elimizde 10 tane konu olduğu için her kelime 0-9 arası konulardan birine atanacaktır. Atama işlemi sonucu Tablo 1’de verilmiştir.

Tablo 1: Rastgele Konu Atama

|  |  |
| --- | --- |
| Doküman No | Konu Atamaları |
| 1 | 3-3-5-3-4-9-1-0-0-6-4-4-5-6-5-8-5-0-4-7-7-4-9-3-6-3-5-5-2-5-1-4-8-5-8-1-9-8 |
| 2 | 3-1-6-3-1-6-8-5-8-3-3-2-4-9-3-5-3-6-6-2-7-1-2-3-8-3-7-5-5 |
| 3 | 6-8-4-7-5 |
| 4 | 4-2-4-6-2-2-2-3 |
| 5 | 4-7-3-1-5-6-9-4-7-9-6-9-0-4-9-1-5-0-7-2-1-4-4-1-6-6-4-3-3-5-4-7-3-9-6-7-4-2-4-8-6-0-5-3-5-6-3-6 |
| 6 | 1-7-5-6-1-6-3-4-0-4-5-8-9-4-5-8-7-7-3-9-2-5-5-0-5-8-9-6-6-5-2-5-4-6-6-9-2-2-5-2-5-9-3-9-6-6-1-7-0-8-6-4-1-3-8-4-4-8-1-8-2-1-6 |
| 7 | 0-8-8-5-7-0-3-1 |
| 8 | 0-6-6-3-6-6-5-1-4-9-1-4-4-9-1-1-2-6-6-9-7-4-2 |
| 9 | 4-3-6-2-6-8-0-9-2-8-0-8-9-6-5-1-8-5-7-9 |
| 10 | 9-4-0-2-1-2-4-7-5-4-3 |

1. Her doküman için konu atama işlemi gerçekleştirildikten sonra bu bilgiyi kullanarak çeşitli istatistikler çıkartılır.
   1. Yerel istatistiklerin yorumlardan çıkartılması: 1. yorumda her konuya kaç adet kelime/kelime öbeği atanmıştır. 1. yorum için elde edilen istatistik Tablo 2’de verilmiştir. Bu işlem tüm koleksiyonda gerçekleştirilir.

Tablo 2: 1. Yorum için Yerel İstatistik

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Konu0 | Konu1 | Konu2 | Konu3 | Konu4 | Konu5 | Konu6 | Konu7 | Konu8 | Konu9 |
| Yorum1 | 3 | 3 | 1 | 5 | 6 | 8 | 3 | 2 | 4 | 3 |

* 1. Global istatistikler koleksiyondan çıkartılır. Yani tüm koleksiyon için her kelimenin her konuya kaç kere atandığı hesaplanır. Örneğin “friend” kelimesi 2, 4, 6 ve 8. konulara birer kez atanmışken geri kalan altı konuya sıfır kere atanmıştır.

1. İstatistiki bilgiler elde edildikten sonra her yorum için her kelimenin yeniden konu ataması gerçekleştirilir. Bunun için mevcut istatistiki bilgilerin güncellenmesi gerekmektedir. Örneğin sadece 1. dokümanda yer alan “cuisine” kelimesi için yerel istatistik değeri Tablo 3’teki gibi güncellenecektir.

Tablo 3: 1. Yorum için Yerel İstatistiğin Güncellenmesi

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Konu0 | Konu1 | Konu2 | **Konu3** | Konu4 | Konu5 | Konu6 | Konu7 | Konu8 | Konu9 |
| Yorum1 | 3 | 3 | 1 | **4** | 6 | 8 | 3 | 2 | 4 | 3 |

1. Yerel istatistikten sonra global istatistik de güncellenecektir ve “cuisine” kelimesi için Konu3 değeri 0 olacaktır.
2. Bu adımda “cuisine” kelimesi yeni bir konuya atanacaktır. Burada iki faktör rol almaktadır. Bunlardan ilki mevcut yorumun konular ile hangi oranlarda “ilişkili” olduğudur. Biz “cuisine” için 1. dokümanın her konu ile ne kadar ilişkili olduğu hesaplanır. Bu hesaplama Eşitlik 1’e göre yapılır.

|  |  |
| --- | --- |
|  | (1) |

1. Buradaki i. yorumda k. konuya atanan kelime sayısını göstermektedir. Yerel istatistik bu adımda kullanılarak değeri  ‘nın belirlenir.  dokümanda yer alan toplam kelime sayısıdır. 1 çıkartılmasının nedeni “cuisine” kelimesinin yok sayılmasıdır.
2. İkinci faktör ise “cusine” kelimesinin konular ile ne kadar “ilişkili olduğunun” hesaplanması ile bulunmaktadır. Bu hesaplama adımında ise global istatistiklerden yararlanış olup kelime verilen konu altında ne kadar kullanılmış bilgisi gerekmektedir. Hesaplama adımı Eşitlik 2’de verilmiştir.

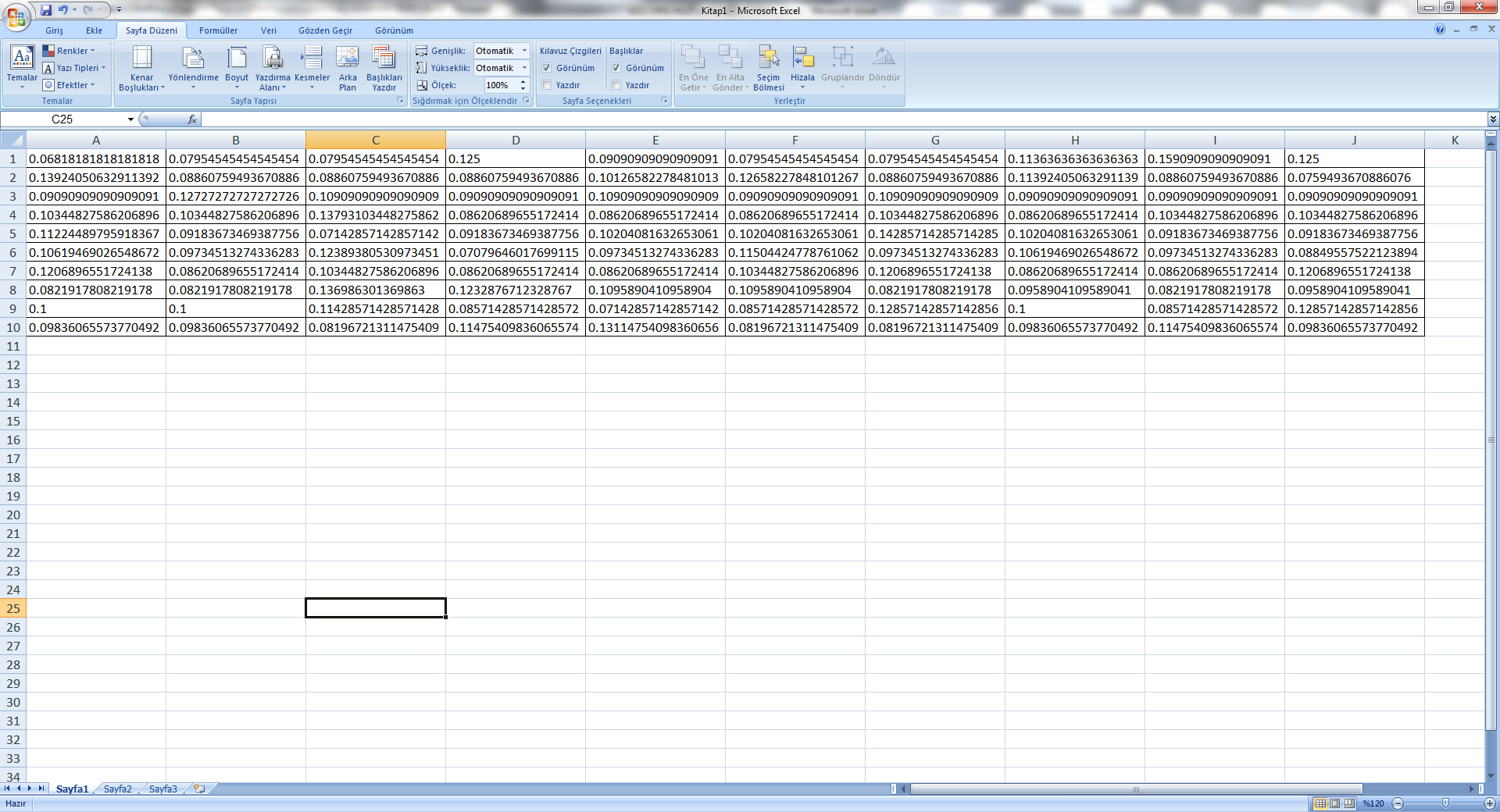
|  |  |
| --- | --- |
|  | (2) |

1.  “cuisine” kelimesinin k. konuya tüm koleksiyonda kaç kere atandığı bilgisini vermektedir. V ise ilk adımda kurulan sözlüğün boyutudur.
2. Eşitlik 1 ve Eşitlik 2’deki değer çarpılarak “cuisine” kelimesinin k. konuya atanma olasılığı hesaplanmaktadır. “cuisine” kelimesi için 7-11 adımları konu sayısı kadar tekrarlanacak ve en yüksek olasılık elde edilen konuya “cuisine” kelimesi için yeni konu olacaktır.
3. Yerel ve global istatistikler güncellenecektir.
4. 5. adımdan itibaren tüm adımlar koleksiyonda yer alan her dokümanın her kelimesi için gerçeklenecek ve yeni konu ataması yapılacaktır.
5. Bu adımda  (konuların yorumlardaki dağılımı ) ve (kelimelerin konulardaki dağılımı)’nın değerleri;  için Eşitlik 1 ve  için Eşitlik 2 kullanılarak gerçeklenecektir. Bu adım ile bir iterasyon tamamlanmış olacaktır. 1. iterasyon sonunda kelimelerin yeni konuları, bu konuların yorumlardaki dağılımı (10x10 matris) ve kelimelerin konulardaki dağılımının (10x187 matris) yeni değerleri elde edilmiş olacaktır.

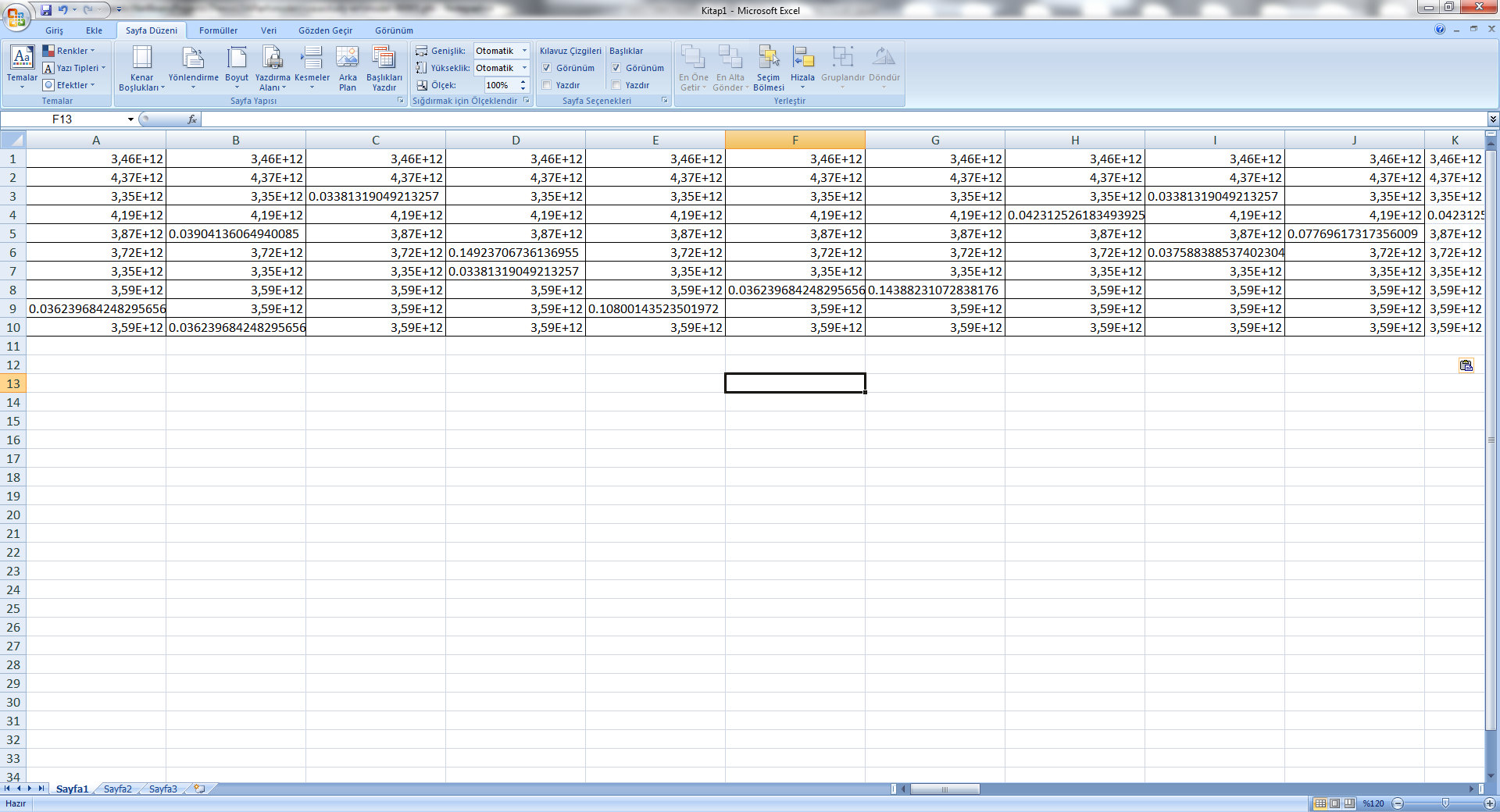
Tablo 4: 1. İterasyon Sonunda Kelimelerin Konuları

|  |  |
| --- | --- |
| Doküman No | Konu Atamaları |
| 1 | 8-9-2-6-8-7-7-3-5-4-3-4-1-6-3-3-4-7-5-9-8-3-8-9-1-9-2-8-8-9-7-7-8-3-8-0-9-8 |
| 2 | 2-0-4-3-1-5-6-7-7-5-2-8-0-0-6-9-4-0-7-4-5-7-8-0-1-5-0-3-5 |
| 3 | 1-2-6-4-1 |
| 4 | 2-6-2-2-1-8-9-0 |
| 5 | 6-6-8-6-8-6-9-6-7-0-5-1-0-7-2-3-7-6-7-9-5-2-9-1-0-0-1-3-0-0-7-1-3-6-9-4-5-4-4-8-8-3-6-4-6-5-5-4 |
| 6 | 1-9-7-7-5-6-0-2-4-6-4-4-4-4-5-8-7-7-2-5-7-8-2-5-2-9-8-8-5-6-2-0-0-2-5-9-4-8-6-2-5-5-6-9-0-8-0-7-0-9-1-1-6-0-2-1-3-2-7-3-1-3-1 |
| 7 | 0-5-2-0-9-6-6-9 |
| 8 | 3-5-1-6-8-9-5-0-2-9-7-4-2-4-3-4-2-3-3-2-5-7-2 |
| 9 | 8-6-6-0-1-2-1-9-6-9-9-6-0-7-5-2-3-7-2-9 |
| 10 | 4-8-3-7-1-4-4-0-8-9-3 |

Tablo 5: 1. İterasyon Sonucunda Elde Edilen Theta Değerleri



Tablo 6: 1. İterasyon Sonucunda Elde Edilen Pi Değerleri

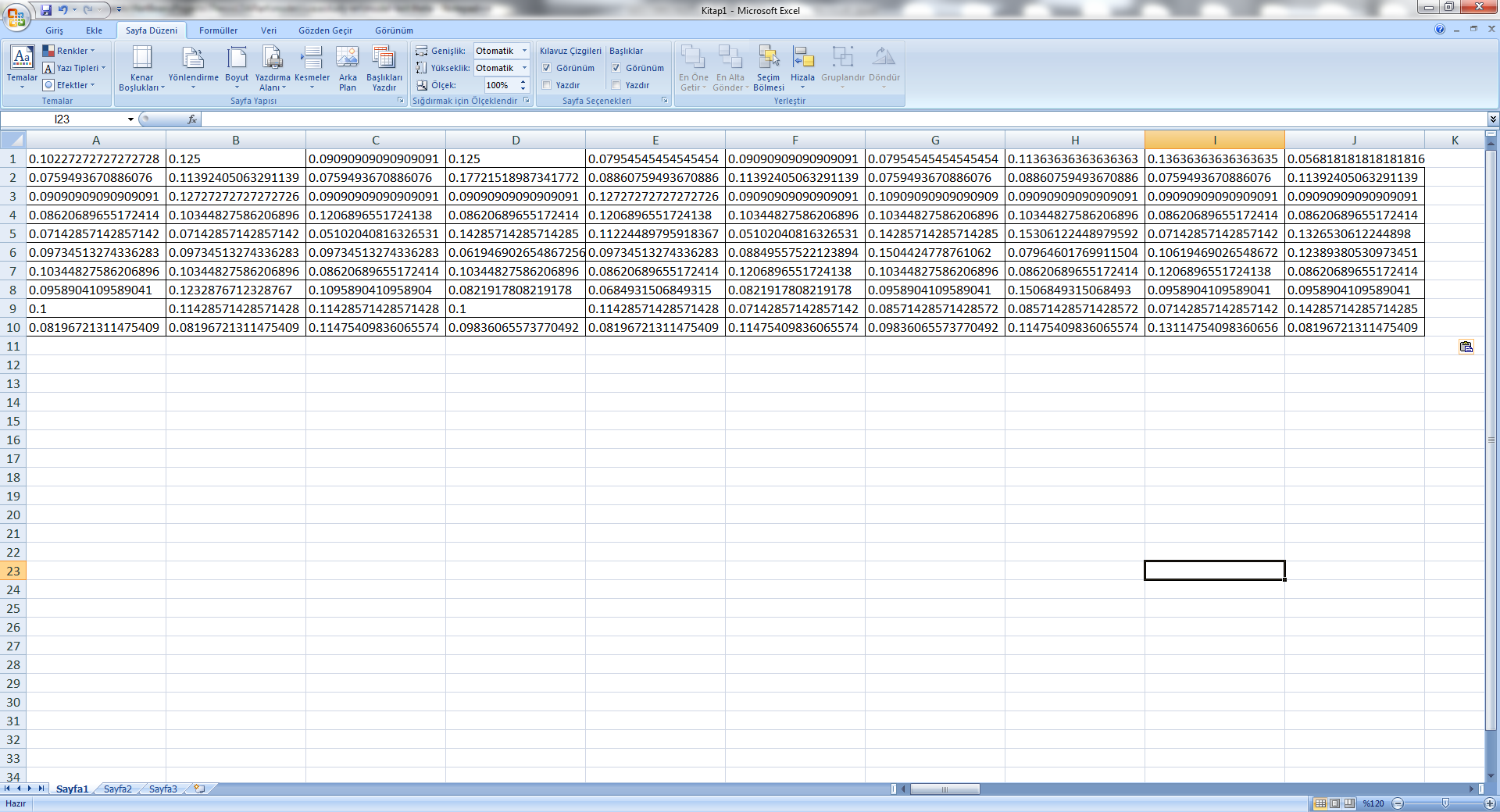


1. 10. iterasyon sonunda elde edilen değerler aşağıdaki tablolarda verilmiştir.

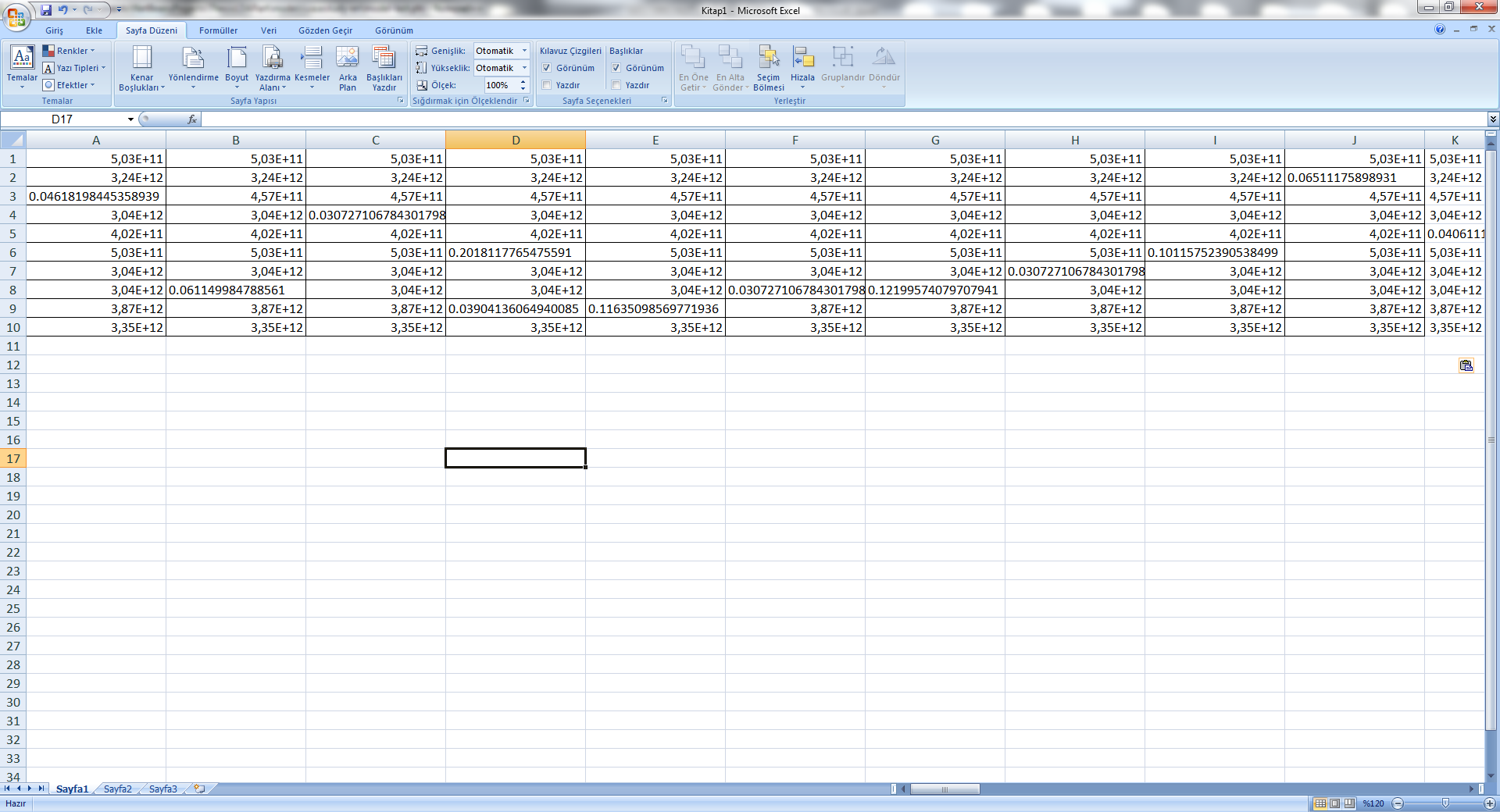
Tablo 7: 10. İterasyon Sonunda Kelimelerin Konuları

|  |  |
| --- | --- |
| Doküman No | Konu Atamaları |
| 1 | 2-7-3-5-8-7-7-6-5-1-4-1-3-8-8-1-0-1-0-8-8-3-3-4-6-5-2-3-7-1-8-2-1-3-8-0-0-7 |
| 2 | 1-3-5-3-7-5-6-3-4-9-5-9-0-3-9-3-3-1-9-5-3-1-8-4-1-3-2-3-7 |
| 3 | 1-4-4-6-1 |
| 4 | 6-4-7-2-1-4-2-5 |
| 5 | 3-6-4-6-9-6-8-7-7-6-8-7-6-3-1-3-7-3-7-7-0-9-9-7-6-7-6-9-9-0-6-7-3-1-3-4-9-6-4-9-3-9-7-3-4-3-4-4 |
| 6 | 1-9-7-0-2-6-0-2-9-0-4-7-1-6-8-8-2-8-2-6-8-4-6-7-5-9-8-1-6-9-2-0-6-0-5-9-5-4-8-6-5-9-1-0-9-3-6-9-4-6-9-3-1-4-8-6-6-2-5-7-4-1-6 |
| 7 | 0-8-5-3-5-6-1-8 |
| 8 | 8-1-6-6-7-7-7-7-1-9-8-7-2-5-0-1-2-9-7-2-1-3-0 |
| 9 | 1-3-4-2-4-2-2-1-9-9-0-6-0-9-7-4-3-1-9-9 |
| 10 | 7-5-5-8-6-7-8-8-2-2-3 |

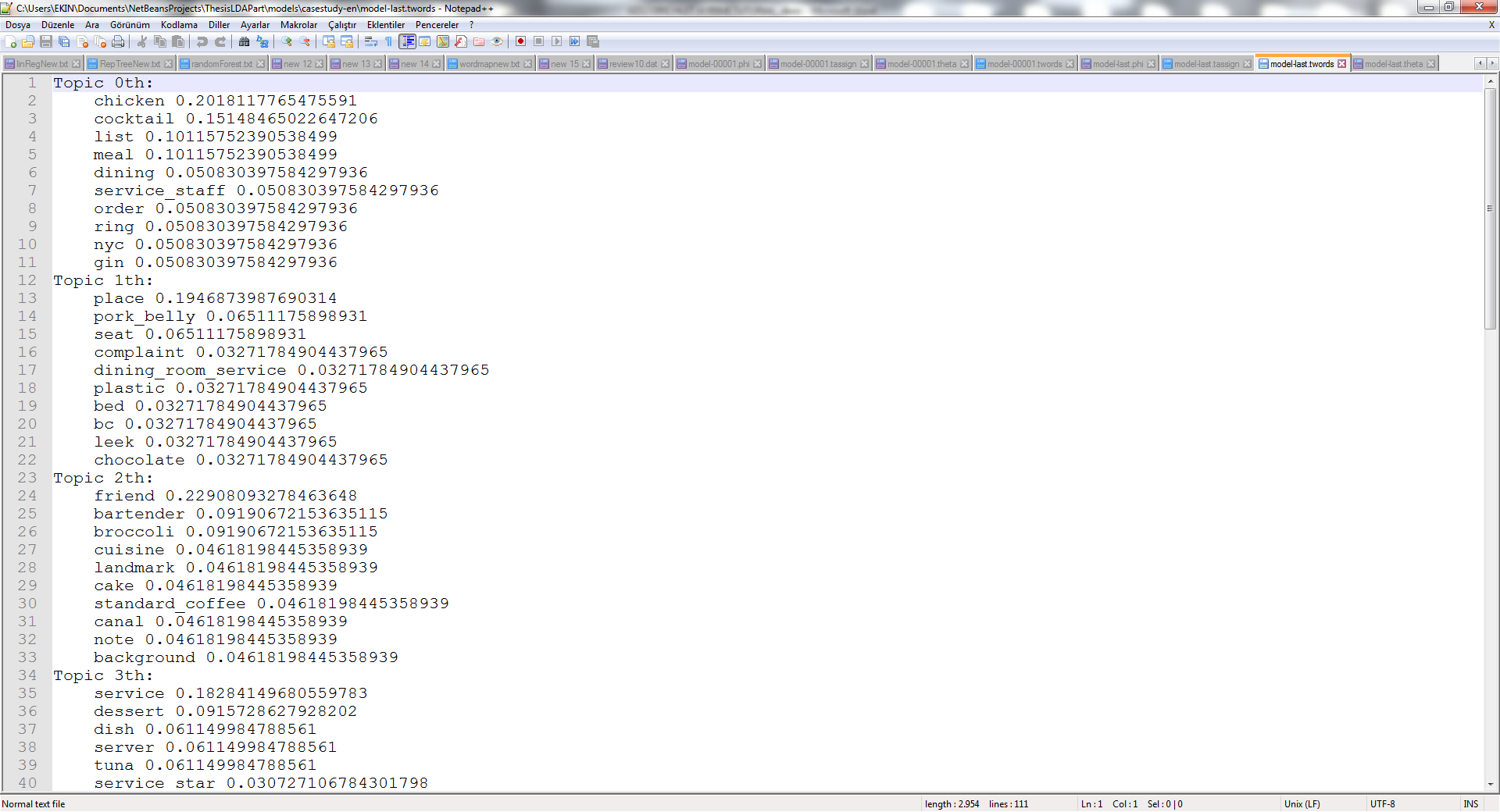
Tablo 8: 10. İterasyon Sonucunda Elde Edilen Theta Değerleri



Tablo 9: 10. İterasyon Sonucunda Elde Edilen Pi Değerleri



1. 10. İterasyonda elde edilen değerleri her konu altında büyükten küçüğe sıralanarak her konu altındaki ilk 10 kelime belirlenmiştir dolayısıyla konular da belirlenmiştir.



**Şekil 2. 10 İterasyon Sonucunda Elde Edilen Topic Wordler**