Human Activity Analysis and Prediction Using Google n-Grams

İlknur Dönmez

***Abstract*—There seems to be no compelling reason to argue that the importance of data is increasing. The huge data from social environments, internet of things and online books can be collected and processed easily via today technology. In social life and business the user and costumer oriented approaches become more important. User behavior, user features and user opinion are searched in different applications. In this paper the human activities are extracted and analyzed using Google-n grams. Google-n grams are generated from millions of books between 1500 to 2000 which can be an indicator for human specific feature and behavior. In this paper human specific main activities are analyzed and the human activities in near feature are predicted via Google n-grams and the functions which are generated via using n-grams.**

***Index Terms*—Google n-grams, activity extraction, activity prediction, human activity mining.**

I. Introduction

"Human language and words express human specific features and behaviors" is the hypothesis in this paper. The main argument that can be advanced to support this hypothesis is the oral and written expressions of languages as an indispensable tool for communication. Language is the main form of expression of though and also opens the way to knowledge, art and culture in general. Writing is a medium of human communication that represents language and emotion with signs and symbols. As a subset of language written expressions, books are expression of though and they also express human knowledge, art and culture. In our paper human activities are analyzed using google n-grams and the relative frequencies of activities are searched for near future.

Books have the knowledge or information that relates to a particular subject, person or anything. Human express believes, needs, hopes, fears and his/herself and the ideas about other people in books. Book involves activities, hobbies and the things that human wonder or give importance. The human can be analyzed with the word and word groups (n-grams) using enough books.

Google n-grams are generated from millions of book between 1500 and 2000. It has a search editor "Google n-gram" viewer to search different word and words group. The x-axis indicates time line and the y-axis indicates the relative frequency, in percentage terms, of the text being searched, over the time period specified. The Google n-Gram Viewer [1] searches and provides results based on a single corpus of books.

Manuscript received March 12, 2018; revised May 5, 2018.

The author is with the Computer Engineering Department, İstanbul Bilgi University in İstanbul, Turkey (e-mail: ilknur.buyukkuscu@hotmail.com).

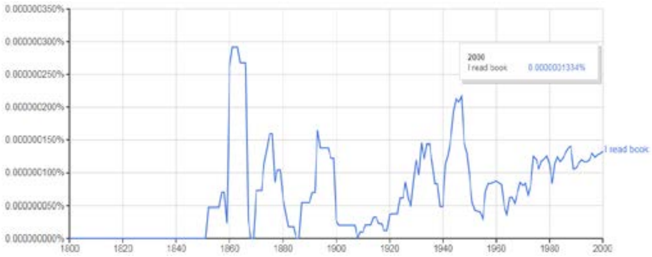


Fig. 1. "I read book "search in Google n-gram viewer.

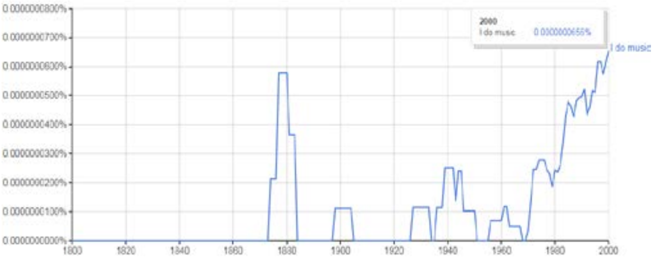


Fig. 2. "I do music" search in Google n-gram viewer.

As seen on the Fig. 1 and Fig. 2 "I read book" and "I do music" are searched in Google n-gram viewer. Maximum five consecutive words can be searched according to desired time period between 1500 and 2000 in Google n-gram viewer.

There are several studies using Google n-grams. These studies are about correlation of term count and document frequency for Google n-grams [2] and expanding word-emotion association lexicon using google n-grams [3]. Google n-grams are used for comparing word relatedness measures [4], text similarity [5] and measuring cultural complexity [6]. Because of google n-grams contain different language datasets; comparison of semantic similarity for different languages [7] is studied using google n-grams. And there is also one more study about making Google Books n-grams useful for a wide range of research on language change [8].

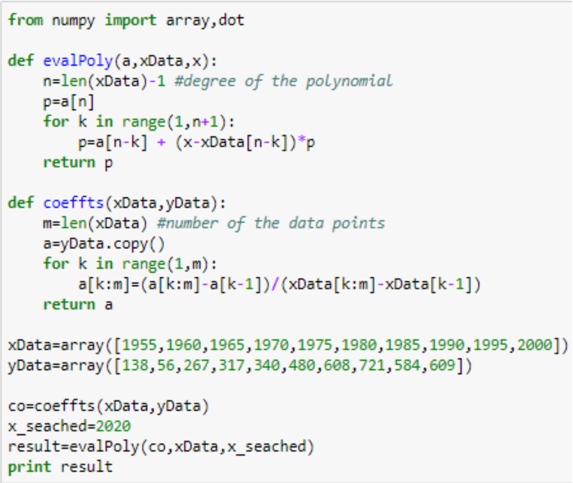
There are also activity selection and extraction studies. A feature extraction method is used for real time human activity recognition on cell phones [9]; there are several studies that searched for human web activities to personalize the Web [10]-[12]. Automatic extraction of human activity knowledge from method-describing web articles [13] and mining models of human activities from the web [14] are studies about extracting human specific activities using web.

Frequency of the occurrence of activity words expressions correlates with the importance of the activity for human. So in this paper, we studied relative frequency of different activities between 1950 and 2000 and predicted the frequency of different activities in 2020.

doi: 10.18178/ijfcc.2018.7.2.516

II. Methods

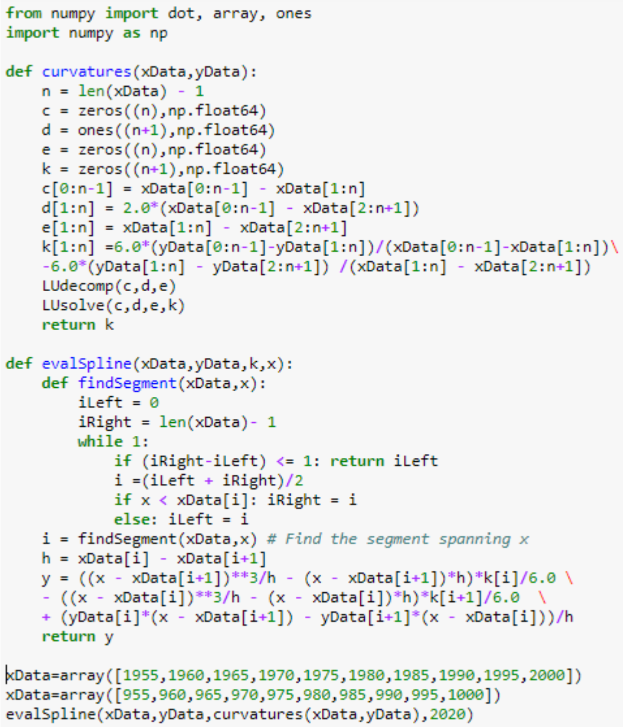
In this paper for expecting the activity frequency according to our data, we used different interpolation and curve fitting techniques in python. Newton's Polynomial Interpolation, Interpolation with Cubic Spline and linear function fitting methods are used. After applying these three methods to the datasets, linear function fitting method has the minimum error rate.



```
from numpy import array,dot

def evalPoly(a,xData,x):
    n=len(xData)-1 #degree of the polynomial
    p=a[n]
    for k in range(1,n+1):
        p=a[n-k] + (x-xData[n-k])*p
    return p

def coeffts(xData,yData):
    m=len(xData) #number of the data points
    a=yData.copy()
    for k in range(1,m):
        a[k:m]=(a[k:m]-a[k-1])/(xData[k:m]-xData[k-1])
    return a

xData=array([1955,1960,1965,1970,1975,1980,1985,1990,1995,2000])
yData=array([138,56,267,317,340,480,608,721,584,609])

co=coeffts(xData,yData)
x_seached=2020
result=evalPoly(co,xData,x_seached)
print result
```

Fig. 3. Newton's Poly. Interpolation for activity frequency expectation.

As seen on the Fig. 3 yData is related with frequency of searched phrase ("I do sport") for the years in xData. In polynomial interpolation, "evalPoly" function evaluates Newton's polynomial p at x. The coefficient vector {a} can be computed by the coefficients function. "coeffts" function computes the coefficients of Newton's polynomial.



```
from numpy import dot, array, ones
import numpy as np

def curvatures(xData,yData):
    n = len(xData) - 1
    c = zeros((n),np.float64)
    d = ones((n+1),np.float64)
    e = zeros((n),np.float64)
    k = zeros((n+1),np.float64)
    c[0:n-1] = xData[0:n-1] - xData[1:n]
    d[1:n] = 2.0*(xData[0:n-1] - xData[2:n+1])
    e[1:n] = xData[1:n] - xData[2:n+1]
    k[1:n] =6.0*(yData[0:n-1]-yData[1:n])/(xData[0:n-1]-xData[1:n])\
    -6.0*(yData[1:n] - yData[2:n+1]) /(xData[1:n] - xData[2:n+1])
    LUdecomp(c,d,e)
    LUsolve(c,d,e,k)
    return k

def evalSpline(xData,yData,k,x):
    def findSegment(xData,x):
        iLeft = 0
        iRight = len(xData)- 1
        while 1:
            if (iRight-iLeft) <= 1: return iLeft
            i =(iLeft + iRight)/2
            if x < xData[i]: iRight = i
            else: iLeft = i
    i = findSegment(xData,x) # Find the segment spanning x
    h = xData[i] - xData[i+1]
    y = ((x - xData[i+1])**3/h - (x - xData[i+1])*h)*k[i]/6.0 \
    - ((x - xData[i])**3/h - (x - xData[i])*h)*k[i+1]/6.0  \
    + (yData[i]*(x - xData[i+1]) - yData[i+1]*(x - xData[i]))/h
    return y

xData=array([1955,1960,1965,1970,1975,1980,1985,1990,1995,2000])
xData=array([955,960,965,970,975,980,985,990,995,1000])
evalSpline(xData,yData,curvatures(xData,yData),2020)
```

Fig. 4. Interpolation with cubic spline for activity frequency expectation.

As seen on the Fig. 4 in Interpolation with cubic spline for activity frequency expectation, yData is related with frequency of searched phrase ("I do sport") for the years in xData. Curvatures function returns the curvatures {k} of cubic spline at the knots. EvalSpline function evaluates cubic spline at x. The curvatures {k} can be computed with the function 'curvatures'.

In linear function fitting method, the linear function that is passing through the weighted average point and the end point is used as seen on Fig. 5.

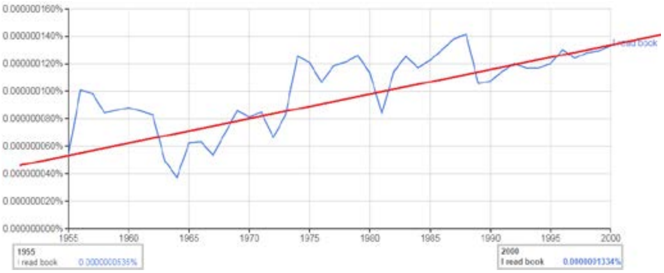


Fig. 5. Linear function estimation for "I read book".

III. Data

Used data is generated from google n-gram datasets. Each frequency is calculated from the occurrence of searched phrase at the related year. So the data shows the frequency of appearing phrase in the books over a period of time. The frequency values for each phrases indicates the relative frequency, in percentage terms, of the text being searched, over the time period specified.

The data seen in Table-I is related with main activities that is encountered via using google n-grams. The searched phrases are "I do music", "I do sport", "I make art" and "I read book" for the main activities.

TABLE I: Total Frequency for Main Activities By Dates

	Phrase Frequency (*10^{-12})			
Date	Do music	Do sport	Make Art	Read book
1955	253	0	0	535
1960	450	69	48	881
1965	352	49	84	624
1970	217	35	389	812
1975	66	279	348	120
1980	129	242	783	1135
1985	48	478	1705	1226
1990	103	495	2023	1075
1995	106	511	2561	1202
2000	145	656	3737	1296

TABLE II: Total Frequency for Art Activities by Dates

	Phrase Frequency (*10^{-12})			
Date	Make pic.	Sculpture	Photo.	Jewelry
1955	1556	0	0	0
1960	1625	0	48	0
1965	607	239	48	41
1970	1238	500	35	70
1975	1534	207	0	374
1980	1743	242	367	217
1985	1820	476	138	180
1990	1636	313	123	266
1995	1721	262	368	505
2000	1646	103	334	566

The data seen in Table II is related with the branches of art activities that are encountered in books via using google n-grams. The searched phrases are "I make picture", "I make sculpture", "I make photography" and "I make jewelry" for the branches of art activities.

IV. Analysis

A. Analyzing Main Activities

First of all via using google-n grams the mostly used activities are selected. In our study these activities are “doing music”, “doing sport”, “making art” and “reading book”. Then we divide making art activity to its sub classes.

When we analyze four main activities using Google n-gram datasets, the average frequencies between 1950 and 2000 of four main activities are shown in Fig.6. There is 10^{-12} coefficient for each frequency. It is found by dividing searched n-gram numbers divided by all n-grams numbers. Art activity has the highest average frequency and the music activity has the smallest average frequency between 1950 and 2000.

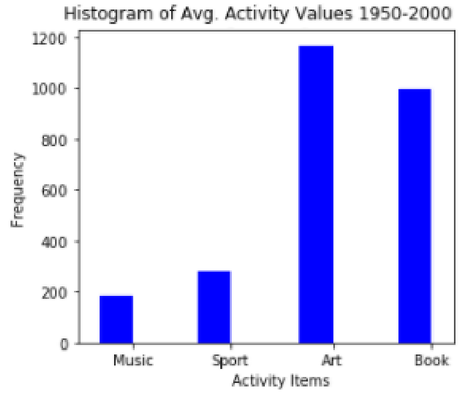


Fig. 6. Histogram of avg. main activity values between 1950 and 2000.

When we look four art activities using Google n-gram datasets, the average frequencies of four art activities are shown in Fig. 7. Picture activity has the highest average frequency and the photography activity has the smallest average frequency between 1950 and 2000.

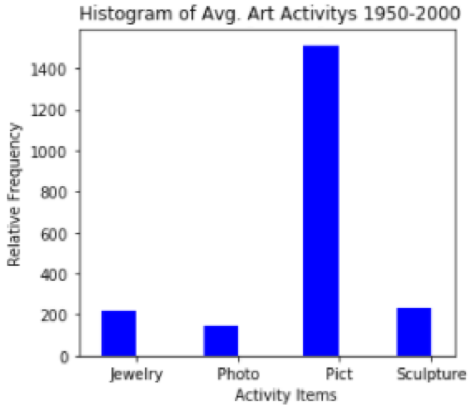


Fig. 7. Histogram of Avg. Art Activities between 1950 and 2000.

B. Predicting Activities in Near Feature

It can be predicted how activity frequency will change via looking at old activity frequency over time. The frequencies of activity samples create discrete data sets. The source of the data is generated using Google n-gram database. Interpolation and curve fitting are used for prediction. In interpolation we construct a curve through the data points. Curve fitting is applied to data that contains scatter, usually due to measurement errors. In curve fitting a smooth curve is found that approximates the data in some sense. Thus the curve does not necessarily hit the data points.

TABLE III: Weights of Linear Functions for Main Activities

Linear Model for	W1	W0
Do music	-1.86	3869.44
Do sport	16.64	-32642.66
Make art	114.18	-224636.33
Read book	13.17	-25059.55

The weights of four linear functions fitting data points of four main activities are seen on Table III.

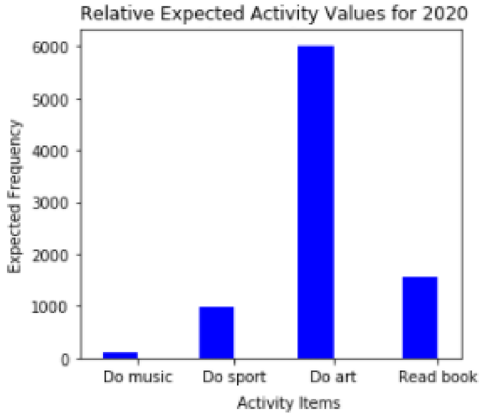


Fig. 8. Histogram of Expected Main Activities on 2020.

In Fig. 8 expected frequencies of four main activities on 2020 are seen. Art activity has the highest expected frequency value and the music activity has the smallest expected frequency value on 2020.

TABLE IV: Weights of Linear Functions for Art Activities

Linear Model for	W1	W0
Make picture	5.90	-10168.22
Make sculpture	-5.38	10885.44
Make photography	8.35	-16369.11
Make jewelry	15.29	-30020.66

The weights of four linear functions fitting data points of four art activities are seen on Table IV.

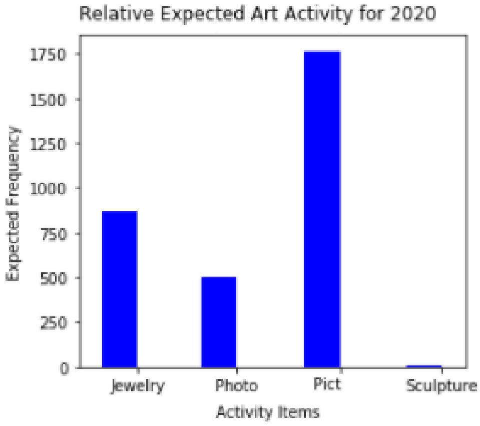


Fig. 9. Histogram of Expected Art Activities on 2020.

In Fig. 9 expected frequencies of four art activities on 2020 are seen. Picture activity has the highest expected frequency value and the sculpture activity has the smallest expected frequency value on 2020.

V. Evaluation

In this paper after applying different methods to the datasets, linear function fitting method has the minimum error rate.

TABLE V: Avg and Y Expected Values for Main Activities

	Do music	Do sport	Make Art	Read book
Avg.	186.9	281.39	1167.8	999.5
y_exp	107.75	998.98	6020.73	1559.5

The average and expected frequencies for main activities are seen on Table V.

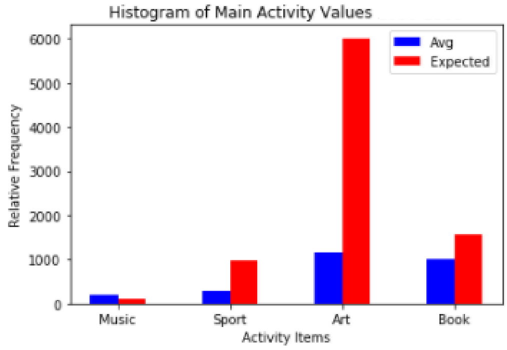


Fig. 10. Histogram of main activities.

As seen on the Fig. 10 the expected frequencies and average frequencies of four main activities are seen on the same scale. The expected value of sport, art and book increase but the expected value of music decrease. The art activities have a huge increment at least four times bigger than the average frequency. Although the average value of doing sport activity is smaller than the reading book activity, nevertheless the increase in doing sport activity is much bigger than the reading book activity.

TABLE VI: Avg and Y Expected Values for Art Activities

	Make pic.	Sculpture	Photo.	Jewelry
Avg.	1513.0	234.19	146.09	221.90
y_exp	1764.14	5.27	501.03	871.86

The average and expected frequencies for art activities are seen on Table VI.

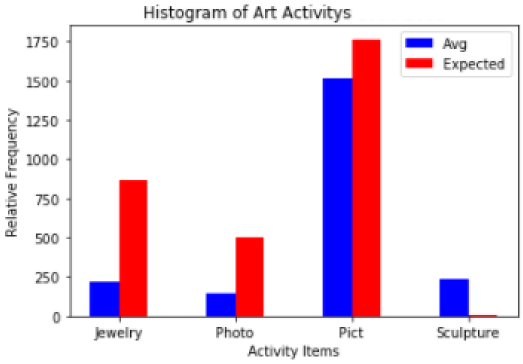


Fig. 11. Histogram of Art Activities

As seen on the Fig.11 the expected frequencies and average frequencies of four art activities are seen on the same scale. The expected values of making jewelry, photography and picture increase but the expected value of making sculpture decreases. The making jewelry activities have a big increment at least two times bigger than the average frequency. Although the average value of making photography activity is smaller than the making photography activity, nevertheless the increase in making photography activity is bigger than the making picture activity. There is a huge decrease in making sculpture activity.

VI. Future Work

Google n-gram datasets provide a very important source for human specific feature and behavior analysis. Our study is an initial work aiming to show what can be done within this field. Using this database it is possible to study, new phrases and concepts that has started to take place in human life and language. For example, using this database we can find that “space colonies” phrase has been used in books since 1955. The “word groups” people have been using to express their feelings and character, and the changes in this “word groups” can be another research topic. It could also be possible to study the adjectives and words used in relation to these concepts and the change of these concepts with time from a human centric viewpoint.

VII. Conclusion

Even the little part of a domain is used; the average mentioned activities and expected changes in these activities can be extracted. The result may be used for companies or government to generate product that reflects the human activity trends, or to revive decreasing activities like making sculpture.

References

[1] A. Weiss, “Google N-Gram viewer,” *The Complete Guide to Using Google in Libraries: Instruction, Administration, and Staff Productivity*, vol. 1, 2015.

[2] M. Klein and M. L. Nelson, “Correlation of term count and document frequency for Google N-Grams,” in *Proc. ECIR*, 2009, vol. 9, pp. 620-627.

[3] J. Perrie, A. Islam, E. Milios, and V. Keselj, “Using google n-grams to expand word-emotion association lexicon,” in *Proc. International Conference on Intelligent Text Processing and Computational Linguistics*, 2013, pp. 137-148.

[4] A. Islam, E. Milios, and V. Keselj, “Comparing word relatedness measures based on Google N-Grams,” in *Proc. COLING 2012: Posters*, 2012, pp. 495-506.

[5] A. Islam, E. Milios, and V. Keselj, “Text similarity using Google Tri-grams,” in *Proc. Canadian Conference on AI*, 2012, vol. 7310, pp. 312-317.

[6] P. Juola, “Using the Google N-Gram corpus to measure cultural complexity,” *Literary and linguistic computing*, *28*(4), 668-675, 2013.

[7] C. Joubarne and D. Inkpen, “Comparison of semantic similarity for different languages using the Google N-Gram corpus and second-order co-occurrence measures,” in *Proc. Canadian Conference on Artificial Intelligence*, 2011, pp. 216-221.

[8] M. Davies, “Making Google Books n-grams useful for a wide range of research on language change,” *International Journal of Corpus Linguistics*, vol. 19, no. 3, pp. 401-416, 2014.

[9] M. Khan, S. I. Ahamed, M. Rahman, and R. O. Smith, “A feature extraction method for realtime human activity recognition on cell phones,” in *Proc. 3rd International Symposium on Quality of Life Technology*, 2011.

[10] F. Radicchi. "Human activity in the web," *Physical Review E*, vol. 80, no. 2, 2009.

[11] R. Barrett, P. P. Maglio, and D. C. Kellem, "How to personalize the Web." in *Proc. the ACM SIGCHI Conference on Human Factors in Computing Systems*, ACM, 1997, pp. 75-82.

[12] M. L. Kasavana, K. Nusair, and K. Teodosic, "Online social networking: Redefining the human web," *Journal of Hospitality and Tourism Technology*, vol. 1, no. 1, pp. 68-82, 2010.

[13] J. Ryu, Y. Jung, K. M. Kim, and S. H. Myaeng. "Automatic extraction of human activity knowledge from method-describing web articles," in *Proc. the 1st Workshop on Automated Knowledge Base Construction*, 2010, p. 16.

[14] M. Perkowitz, M.Philipose, K. Fishkin, and D. J. Patterson, "Mining models of human activities from the web," in *Proc. the 13th International Conference on World Wide Web*, ACM, 2004, pp. 573-582.

İ. Dönmez works as a lecturer in the Department of Computer Engineering, Bilgi University, İstanbul. She has a bachelor degree in Electronic and Communication Department from İstanbul Technical University, has a master degree from Mathematic Department in Ege University and doctorate degree from Computer Engineering Department, İstanbul Technical University. Her major field of study is data science, machine learning and natural language processing. She has more than 10 years' work experience as senior software developer and project manager in industry. Most recently, she has conducted many studies on semantic matrix representation of Turkish sentences, application of semantic in QA and document classification, best context free grammar representation of Turkish.