Sentiment Analysis on Consitutions

By **Athina Panotopoulou**[1]

[1]Dartmouth College, Department of Computer Science

labMT

labMT

- **labMT** 1.0: a data set of 10222 ranked words.

labMT

- **labMT** 1.0: a data set of 10222 ranked words.
- Union of 4 sets (10222):

labMT

- **labMT** 1.0: a data set of 10222 ranked words.
- Union of 4 sets (10222):
  1. 5000 most frequent words in **Twitter**

labMT

- **labMT** 1.0: a data set of 10222 ranked words.
- Union of 4 sets (10222):
  1. 5000 most frequent words in **Twitter**
  2. 5000 most frequent words in **Google Books**

labMT

- **labMT** 1.0: a data set of 10222 ranked words.
- Union of 4 sets (10222):
  1. 5000 most frequent words in **Twitter**
  2. 5000 most frequent words in **Google Books**
  3. 5000 most frequent words in **music lyrics**

labMT

- **labMT** 1.0: a data set of 10222 ranked words.
- Union of 4 sets (10222):
  1. 5000 most frequent words in **Twitter**
  2. 5000 most frequent words in **Google Books**
  3. 5000 most frequent words in **music lyrics**
  4. 5000 most frequent words in **New York Times**

labMT

- **labMT** 1.0: a data set of 10222 ranked words.
- Union of 4 sets (10222):
  1. 5000 most frequent words in **Twitter**
  2. 5000 most frequent words in **Google Books**
  3. 5000 most frequent words in **music lyrics**
  4. 5000 most frequent words in **New York Times**
- The ranking of these words obtained from humans using **Amazon's Mechanical Turk**.

labMT

- **labMT** 1.0: a data set of 10222 ranked words.
- Union of 4 sets (10222):
  1. 5000 most frequent words in **Twitter**
  2. 5000 most frequent words in **Google Books**
  3. 5000 most frequent words in **music lyrics**
  4. 5000 most frequent words in **New York Times**
- The ranking of these words obtained from humans using **Amazon's Mechanical Turk**.
- The ranking is from 1 to 9: 1 **SAD**, and 9 **HAPPY**.

labMT

- **labMT** 1.0: a data set of 10222 ranked words.
- Union of 4 sets (10222):
  1. 5000 most frequent words in **Twitter**
  2. 5000 most frequent words in **Google Books**
  3. 5000 most frequent words in **music lyrics**
  4. 5000 most frequent words in **New York Times**
- The ranking of these words obtained from humans using **Amazon's Mechanical Turk**.
- The ranking is from 1 to 9: 1 **SAD**, and 9 **HAPPY**.
- The ranking is the average of all rankings.

labMT

- **labMT** 1.0: a data set of 10222 ranked words.
- Union of 4 sets (10222):
  1. 5000 most frequent words in **Twitter**
  2. 5000 most frequent words in **Google Books**
  3. 5000 most frequent words in **music lyrics**
  4. 5000 most frequent words in **New York Times**
- The ranking of these words obtained from humans using **Amazon's Mechanical Turk**.
- The ranking is from 1 to 9: 1 **SAD**, and 9 **HAPPY**.
- The ranking is the average of all rankings.

No information about the number of the persons that ranked.

labMT

- **labMT** 1.0: a data set of 10222 ranked words.
- Union of 4 sets (10222):
  1. 5000 most frequent words in **Twitter**
  2. 5000 most frequent words in **Google Books**
  3. 5000 most frequent words in **music lyrics**
  4. 5000 most frequent words in **New York Times**
- The ranking of these words obtained from humans using **Amazon's Mechanical Turk**.
- The ranking is from 1 to 9: 1 **SAD**, and 9 **HAPPY**.
- The ranking is the average of all rankings.

No information about the number of the persons that ranked.
No information about the nationality of the persons that ranked.

labMT

- **labMT** 1.0: a data set of 10222 ranked words.
- Union of 4 sets (10222):
  1. 5000 most frequent words in **Twitter**
  2. 5000 most frequent words in **Google Books**
  3. 5000 most frequent words in **music lyrics**
  4. 5000 most frequent words in **New York Times**
- The ranking of these words obtained from humans using **Amazon's Mechanical Turk**.
- The ranking is from 1 to 9: 1 **SAD**, and 9 **HAPPY**.
- The ranking is the average of all rankings.

No information about the number of the persons that ranked.
No information about the nationality of the persons that ranked.

Definition

We denote with **h(w)** the estimate of average **happiness** for each word **w** $\in$ *labMT*.

Using subset of the initial word list.

Using subset of the initial word list.

Using subset of the initial word list.

- Exclude words that their ranking is between $5 - \Delta H < h(w) < \Delta h + 5$.

Using subset of the initial word list.

- Exclude words that their ranking is between $5 - \Delta H < h(w) < \Delta h + 5$.
  Remove neutral words, to enhance differences!

Using subset of the initial word list.

- Exclude words that their ranking is between $5 - \Delta H < h(w) < \Delta h + 5$.
  Remove neutral words, to enhance differences!
- $\Delta H = 1$ Number of words: 3731

Using subset of the initial word list.

- Exclude words that their ranking is between $5 - \Delta H < h(w) < \Delta h + 5$.
  Remove neutral words, to enhance differences!
- $\Delta H = 1$ Number of words: 3731
- $\Delta H = 2$ Number of words: 1008

Using subset of the initial word list.

- Exclude words that their ranking is between $5 - \Delta H < h(w) < \Delta h + 5$.
  Remove neutral words, to enhance differences!
- $\Delta H = 1$ Number of words: 3731
- $\Delta H = 2$ Number of words: 1008
- $\Delta H = 3$ Number of words: 77

Using subset of the initial word list.

- Exclude words that their ranking is between $5 - \Delta H < h(w) < \Delta h + 5$.
  Remove neutral words, to enhance differences!
- $\Delta H = 1$ Number of words: 3731
- $\Delta H = 2$ Number of words: 1008
- $\Delta H = 3$ Number of words: 77
- Using different subsets of labMT highlight different aspects of our data.

Using subset of the initial word list.

- Exclude words that their ranking is between $5 - \Delta H < h(w) < \Delta h + 5$.
  Remove neutral words, to enhance differences!
- $\Delta H = 1$ Number of words: 3731
- $\Delta H = 2$ Number of words: 1008
- $\Delta H = 3$ Number of words: 77
- Using different subsets of labMT highlight different aspects of our data.

Example

Use words the have happiness ranking between 7 and 9, highlights the positive aspect of a text.

The Dataset

The Dataset

Definition

Constitutions: Our data set consists of 104 **constitutions** from 89 countries.

The Dataset

Definition

Constitutions: Our data set consists of 104 **constitutions** from 89 countries.

Every file of the data set is a *.txt* file.

The Dataset

Definition

Constitutions: Our data set consists of 104 **constitutions** from 89 countries.

Every file of the data set is a *.txt* file.

The file name is related to the **country** and the **date**.

The Dataset

Definition

Constitutions: Our data set consists of 104 **constitutions** from 89 countries.
Every file of the data set is a *.txt* file.
The file name is related to the **country** and the **date**.

Constitutions: $14 \in [1787 - 1898]$; $46 \in [1904 - 1999]$; $44 \in [2000 - 2008]$

The Dataset

Definition

Constitutions: Our data set consists of 104 **constitutions** from 89 countries.
Every file of the data set is a *.txt* file.
The file name is related to the **country** and the **date**.

Constitutions: $14 \in [1787 - 1898]$; $46 \in [1904 - 1999]$; $44 \in [2000 - 2008]$

- Countries that do not exist.

The Dataset

Definition

Constitutions: Our data set consists of 104 **constitutions** from 89 countries.
Every file of the data set is a *.txt* file.
The file name is related to the **country** and the **date**.

Constitutions: $14 \in [1787 - 1898]$; $46 \in [1904 - 1999]$; $44 \in [2000 - 2008]$

- Countries that do not exist.
- Vocabulary that is different from today.

The Dataset

Definition

Constitutions: Our data set consists of 104 **constitutions** from 89 countries.
Every file of the data set is a *.txt* file.
The file name is related to the **country** and the **date**.

Constitutions: $14 \in [1787 - 1898]$; $46 \in [1904 - 1999]$; $44 \in [2000 - 2008]$

- Countries that do not exist.
- Vocabulary that is different from today.
- Every file is written in English, who translated the files?

The Dataset

Definition

Constitutions: Our data set consists of 104 **constitutions** from 89 countries.
Every file of the data set is a *.txt* file.
The file name is related to the **country** and the **date**.

Constitutions: $14 \in [1787 - 1898]$; $46 \in [1904 - 1999]$; $44 \in [2000 - 2008]$

- Countries that do not exist.
- Vocabulary that is different from today.
- Every file is written in English, who translated the files?
- Translation cannot fully transfer the emotions that has the initial word.

The Dataset

Definition

Constitutions: Our data set consists of 104 **constitutions** from 89 countries.
Every file of the data set is a *.txt* file.
The file name is related to the **country** and the **date**.

Constitutions: $14 \in [1787 - 1898]$; $46 \in [1904 - 1999]$; $44 \in [2000 - 2008]$

- Countries that do not exist.
- Vocabulary that is different from today.
- Every file is written in English, who translated the files?
- Translation cannot fully transfer the emotions that has the initial word.

Example

The word **America** has different mood for Bosnia and America.

The Dataset

Definition

Constitutions: Our data set consists of 104 **constitutions** from 89 countries.
Every file of the data set is a *.txt* file.
The file name is related to the **country** and the **date**.

Constitutions: $14 \in [1787 - 1898]$; $46 \in [1904 - 1999]$; $44 \in [2000 - 2008]$

- Countries that do not exist.
- Vocabulary that is different from today.
- Every file is written in English, who translated the files?
- Translation cannot fully transfer the emotions that has the initial word.

Example

The word **America** has different mood for Bosnia and America.The word **liberal** has different meaning for America and Greece.

The Dataset

Definition

Constitutions: Our data set consists of 104 **constitutions** from 89 countries.
Every file of the data set is a *.txt* file.
The file name is related to the **country** and the **date**.

Constitutions: $14 \in [1787 - 1898]$; $46 \in [1904 - 1999]$; $44 \in [2000 - 2008]$

- Countries that do not exist.
- Vocabulary that is different from today.
- Every file is written in English, who translated the files?
- Translation cannot fully transfer the emotions that has the initial word.

Example

The word **America** has different mood for Bosnia and America.The word **liberal** has different meaning for America and Greece.

The preprocessing is not a difficult procedure here, because we are searching for the **exact word**:

The preprocessing is not a difficult procedure here, because we are searching for the **exact word**:

Example

"we've", "you've": two distinct words

1. Convert all characters to lower case.

The preprocessing is not a difficult procedure here, because we are searching for the **exact word**:

Example

"we've", "you've": two distinct words

1. Convert all characters to lower case.
2. Remove special characters such as : .,?-:

The preprocessing is not a difficult procedure here, because we are searching for the **exact word**:

Example

"we've", "you've": two distinct words

1. Convert all characters to lower case.
2. Remove special characters such as : .,?-:
3. Replace with gaps.

The Algorithm

The Algorithm

-Load the labMT.

The Algorithm

-Load the labMT.

-For each Constitution c:

The Algorithm

-Load the labMT.

-For each Constitution c:

1. Create the set of words C that are in the constitution c

The Algorithm

-Load the labMT.

-For each Constitution c:

1. Create the set of words C that are in the constitution c
2. Compute the frequency $f(w)$, for each word w in C.

The Algorithm

-Load the labMT.

-For each Constitution c:

1. Create the set of words C that are in the constitution c
2. Compute the frequency $f(w)$, for each word w in C.
3. We define N as the set of words that are both in C and in labMT: $N = C \bigcap labMT$.

The Algorithm

-Load the labMT.

-For each Constitution c:

1. Create the set of words C that are in the constitution c
2. Compute the frequency $f(w)$, for each word w in C.
3. We define N as the set of words that are both in C and in labMT: $N = C \bigcap labMT$.
4. For each word w in N we have a rank $h(w)$.

The Algorithm

-Load the labMT.

-For each Constitution c:

1. Create the set of words C that are in the constitution c
2. Compute the frequency $f(w)$, for each word w in C.
3. We define N as the set of words that are both in C and in labMT: $N = C \bigcap labMT$.
4. For each word w in N we have a rank $h(w)$.
5. The ranking of the constitution c can then computed by:

The Algorithm

-Load the labMT.

-For each Constitution c:

1. Create the set of words C that are in the constitution c
2. Compute the frequency $f(w)$, for each word w in C.
3. We define N as the set of words that are both in C and in labMT: $N = C \bigcap labMT$.
4. For each word w in N we have a rank $h(w)$.
5. The ranking of the constitution c can then computed by: $h_{avg(c)} = \frac{\sum_{w \in N} h(w) f(w)}{\sum_{w \in N} f(w)}$

The Algorithm

-Load the labMT.

-For each Constitution c:

1. Create the set of words C that are in the constitution c
2. Compute the frequency $f(w)$, for each word w in C.
3. We define N as the set of words that are both in C and in labMT: $N = C \bigcap labMT$.
4. For each word w in N we have a rank $h(w)$.
5. The ranking of the constitution c can then computed by: $h_{avg(c)} = \frac{\sum_{w \in N} h(w) f(w)}{\sum_{w \in N} f(w)}$

We denote by $h_{avg(c)}$ the **happiness ranking of each constitution**.

The results

The results

For each $\Delta h = \{0, 1, 2, 3\}$:

The results

For each $\Delta h = \{0, 1, 2, 3\}$:

- An **.xls file** with the rankings for all constitutions:

The results

For each $\Delta h = \{0, 1, 2, 3\}$:

- An **.xls file** with the rankings for all constitutions:
  The first column is the name of the country, the second the year, and the third the $h_{avg(c)}$.

The results

For each $\Delta h = \{0, 1, 2, 3\}$:

- An **.xls file** with the rankings for all constitutions:
  The first column is the name of the country, the second the year, and the third the $h_{avg(c)}$.
- **Two Histograms**:

The results

For each $\Delta h = \{0, 1, 2, 3\}$:

- An **.xls file** with the rankings for all constitutions:
  The first column is the name of the country, the second the year, and the third the $h_{avg(c)}$.
- **Two Histograms**:
  Complete range of h_{avg} from 1 to 9

The results

For each $\Delta h = \{0, 1, 2, 3\}$:

- An **.xls file** with the rankings for all constitutions:
  The first column is the name of the country, the second the year, and the third the $h_{avg(c)}$.
- **Two Histograms**:
  Complete range of h_{avg} from 1 to 9
  For the specific Δh from min to max approximately

The results

For each $\Delta h = \{0, 1, 2, 3\}$:

- An **.xls file** with the rankings for all constitutions:
  The first column is the name of the country, the second the year, and the third the $h_{avg(c)}$.
- **Two Histograms**:
  Complete range of h_{avg} from 1 to 9
  For the specific Δh from min to max approximately
  x_{axes} = happiness ranking, $y_{axes} \propto$number of constitutions.

The results

For each $\Delta h = \{0, 1, 2, 3\}$:

- An **.xls file** with the rankings for all constitutions:
  The first column is the name of the country, the second the year, and the third the $h_{avg(c)}$.
- **Two Histograms**:
  Complete range of h_{avg} from 1 to 9
  For the specific Δh from min to max approximately
  x_{axes} = happiness ranking, $y_{axes} \propto$number of constitutions.
- **Heatmap**:

The results

For each $\Delta h = \{0, 1, 2, 3\}$:

- An **.xls file** with the rankings for all constitutions:
  The first column is the name of the country, the second the year, and the third the $h_{avg(c)}$.
- **Two Histograms**:
  Complete range of h_{avg} from 1 to 9
  For the specific Δh from min to max approximately
  x_{axes} = happiness ranking, $y_{axes} \propto$number of constitutions.
- **Heatmap**:
  For each country the ranking of the most recent document.

The results

For each $\Delta h = \{0, 1, 2, 3\}$:

- An **.xls file** with the rankings for all constitutions:
  The first column is the name of the country, the second the year, and the third the $h_{avg(c)}$.
- **Two Histograms**:
  Complete range of h_{avg} from 1 to 9
  For the specific Δh from min to max approximately
  x_{axes} = happiness ranking, $y_{axes} \propto$number of constitutions.
- **Heatmap**:
  For each country the ranking of the most recent document.
  Some of the countries do not exist anymore.

The results

For each $\Delta h = \{0, 1, 2, 3\}$:

- An **.xls file** with the rankings for all constitutions:
  The first column is the name of the country, the second the year, and the third the $h_{avg(c)}$.
- **Two Histograms**:
  Complete range of h_{avg} from 1 to 9
  For the specific Δh from min to max approximately
  x_{axes} = happiness ranking, $y_{axes} \propto$number of constitutions.
- **Heatmap**:
  For each country the ranking of the most recent document.
  Some of the countries do not exist anymore.
  Some of the countries are two small to see on the map.

The results

For each $\Delta h = \{0, 1, 2, 3\}$:

- An **.xls file** with the rankings for all constitutions:
  The first column is the name of the country, the second the year, and the third the $h_{avg(c)}$.
- **Two Histograms**:
  Complete range of h_{avg} from 1 to 9
  For the specific Δh from min to max approximately
  x_{axes} = happiness ranking, $y_{axes} \propto$number of constitutions.
- **Heatmap**:
  For each country the ranking of the most recent document.
  Some of the countries do not exist anymore.
  Some of the countries are two small to see on the map.
  Rename of some files to be in accordance with the current country naming.

The results

For each $\Delta h = \{0, 1, 2, 3\}$:

- An **.xls file** with the rankings for all constitutions:
  The first column is the name of the country, the second the year, and the third the $h_{avg(c)}$.
- **Two Histograms**:
  Complete range of h_{avg} from 1 to 9
  For the specific Δh from min to max approximately
  x_{axes} = happiness ranking, $y_{axes} \propto$number of constitutions.
- **Heatmap**:
  For each country the ranking of the most recent document.
  Some of the countries do not exist anymore.
  Some of the countries are two small to see on the map.
  Rename of some files to be in accordance with the current country naming.

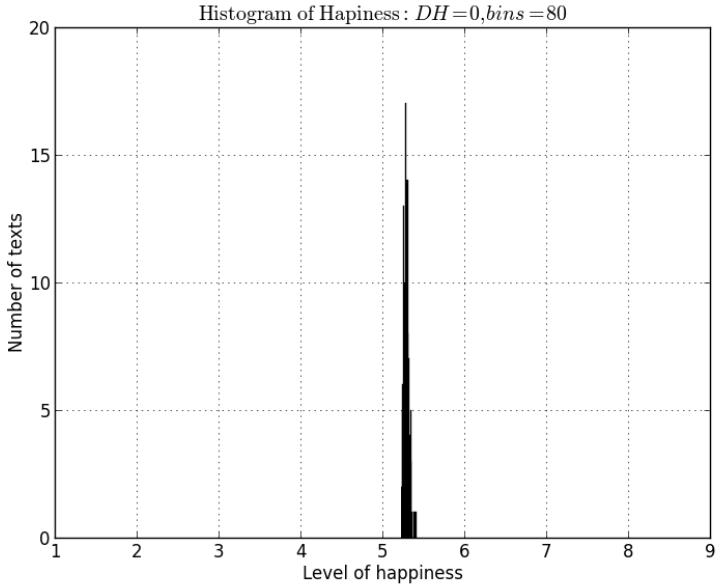
Some of the files have not a ranking for $\Delta h = 3$

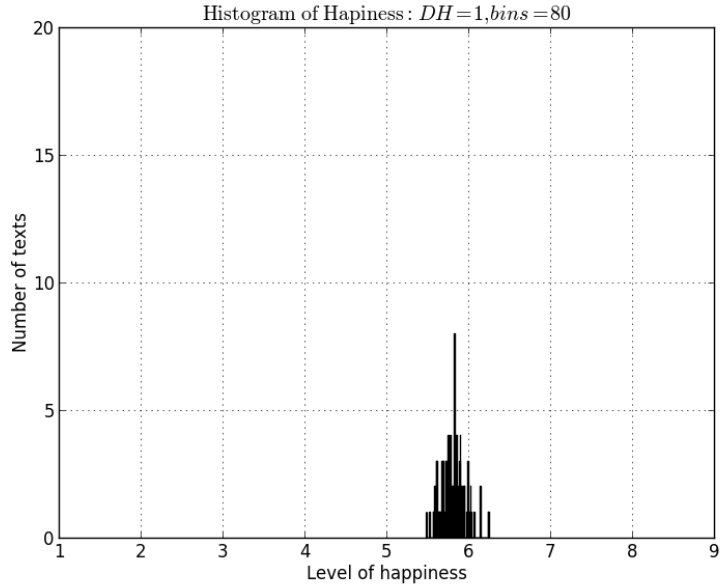
The results

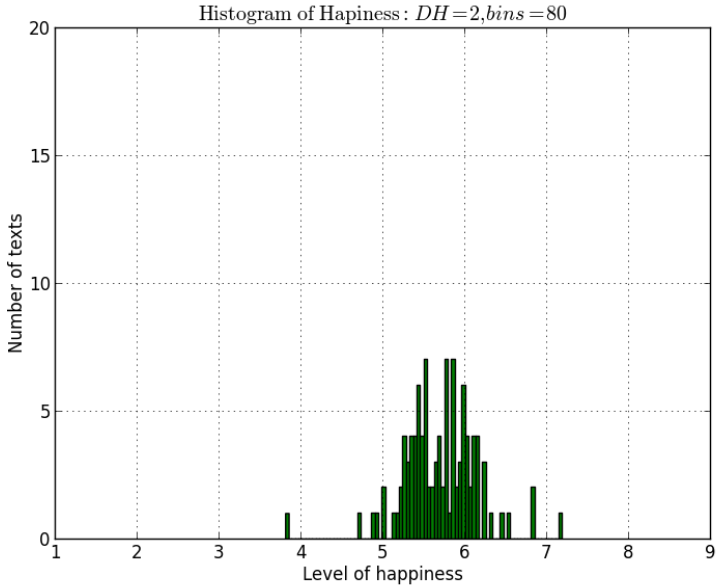
For each $\Delta h = \{0, 1, 2, 3\}$:

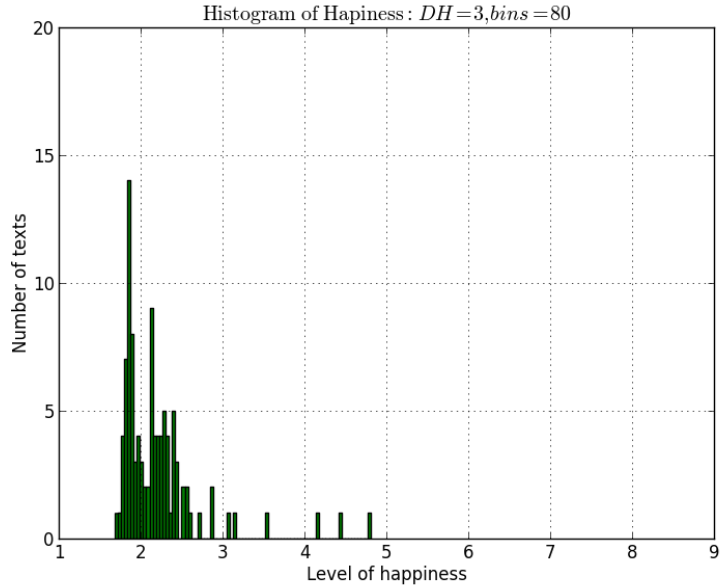
- An **.xls file** with the rankings for all constitutions:
  The first column is the name of the country, the second the year, and the third the $h_{avg(c)}$.
- **Two Histograms**:
  Complete range of h_{avg} from 1 to 9
  For the specific Δh from min to max approximately
  x_{axes} = happiness ranking, $y_{axes} \propto$number of constitutions.
- **Heatmap**:
  For each country the ranking of the most recent document.
  Some of the countries do not exist anymore.
  Some of the countries are two small to see on the map.
  Rename of some files to be in accordance with the current country naming.

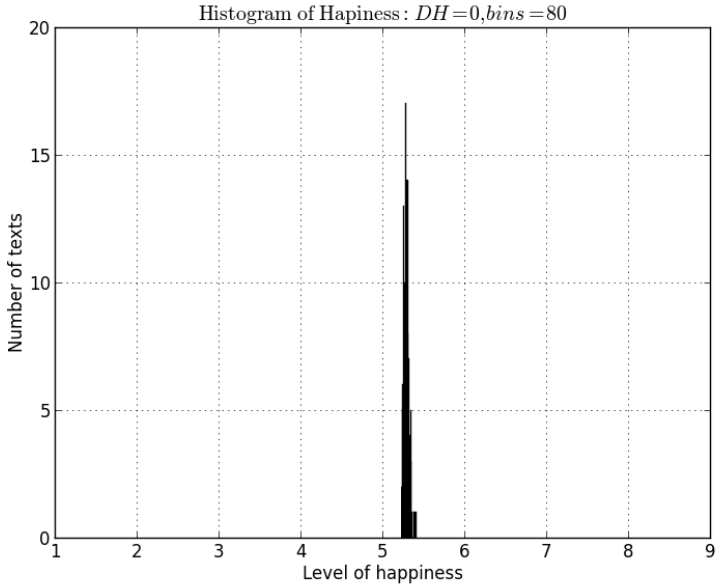
Some of the files have not a ranking for $\Delta h = 3$

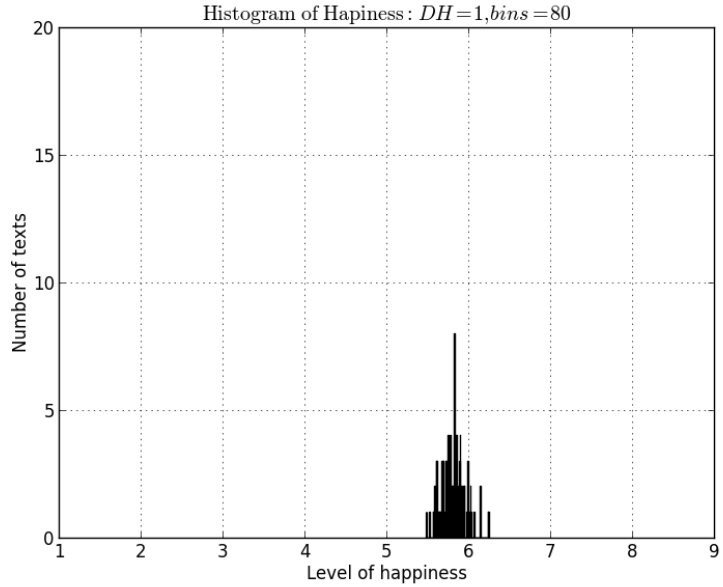

Histogram of Hapiness : $DH = 0, bins = 80$
Number of texts
Level of happiness
0
5
10
15
20
1
2
3
4
5
6
7
8
9

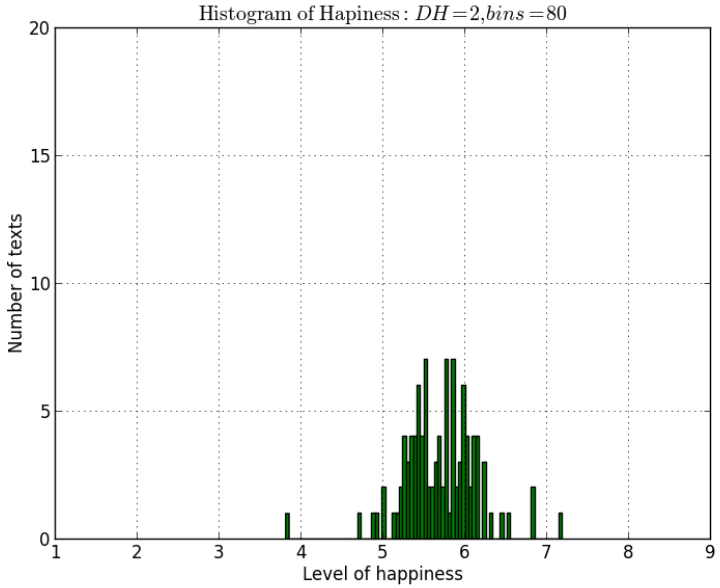

Histogram of Hapiness : $DH = 1, bins = 80$
Number of texts
Level of happiness
0
5
10
15
20
1
2
3
4
5
6
7
8
9

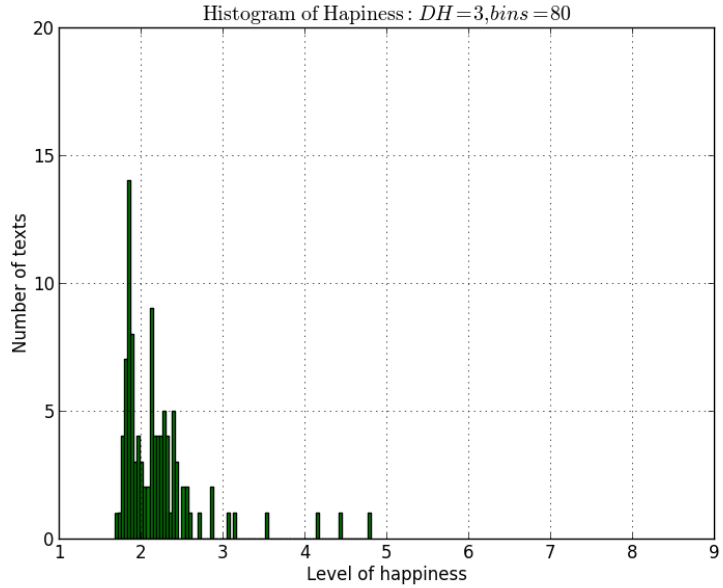

Histogram of Hapiness : $DH = 2, bins = 80$
Number of texts
Level of happiness
0
5
10
15
20
1
2
3
4
5
6
7
8
9

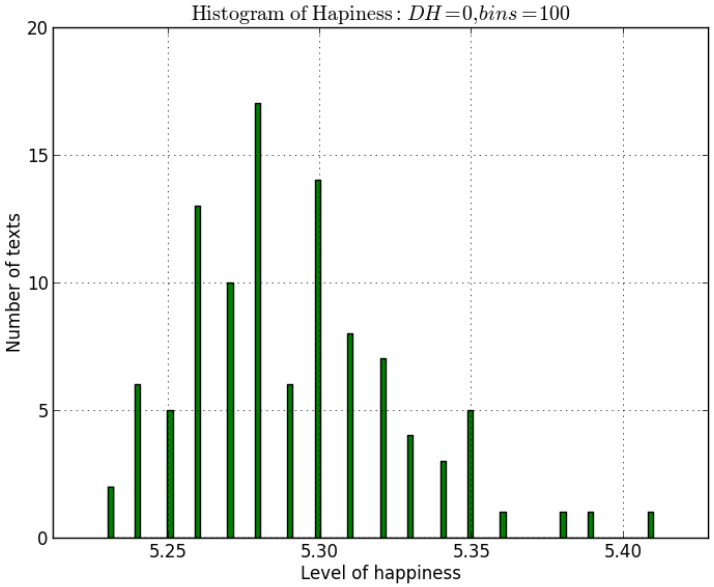

Histogram of Hapiness : $DH = 3, bins = 80$
Number of texts
Level of happiness
0
5
10
15
20
1
2
3
4
5
6
7
8
9

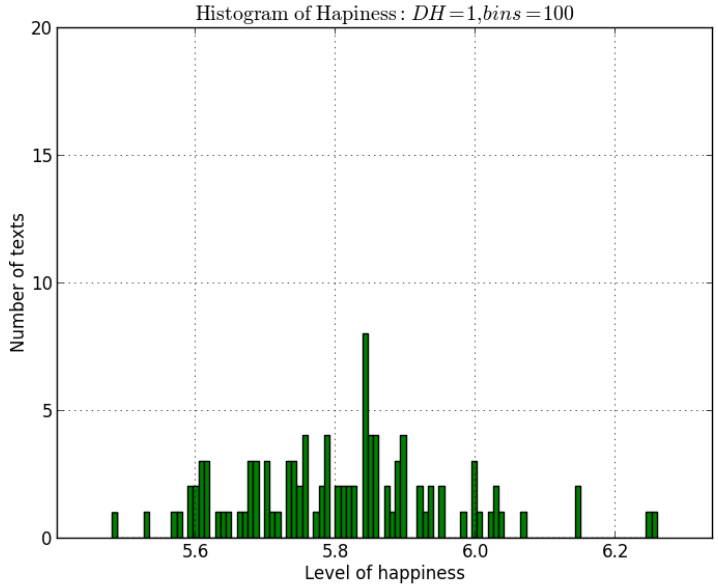

Histogram of Hapiness : $DH = 0, bins = 80$
Number of texts
Level of happiness
0
5
10
15
20
1
2
3
4
5
6
7
8
9

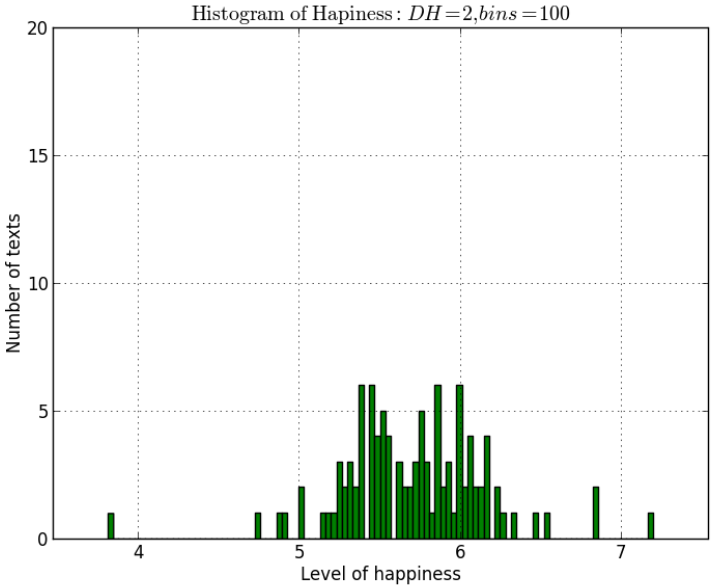

Histogram of Hapiness : $DH = 1, bins = 80$
Number of texts
Level of happiness
0
5
10
15
20
1
2
3
4
5
6
7
8
9

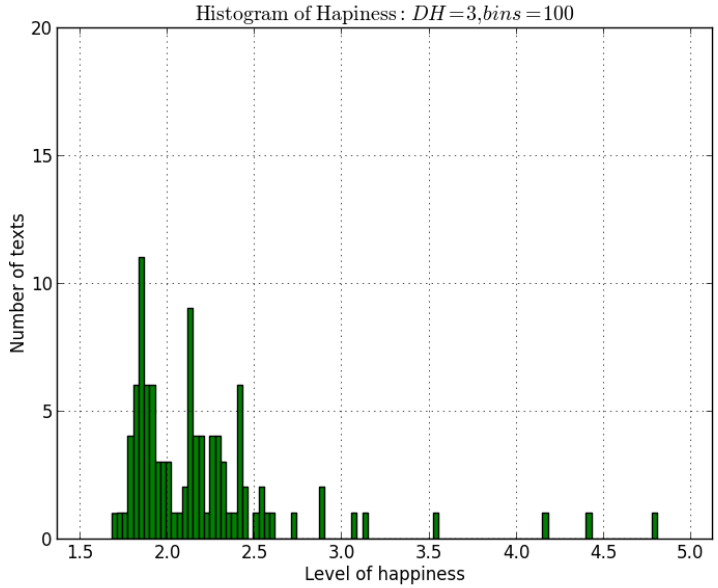

Histogram of Hapiness : $DH = 2, bins = 80$
Number of texts
Level of happiness
0
5
10
15
20
1
2
3
4
5
6
7
8
9

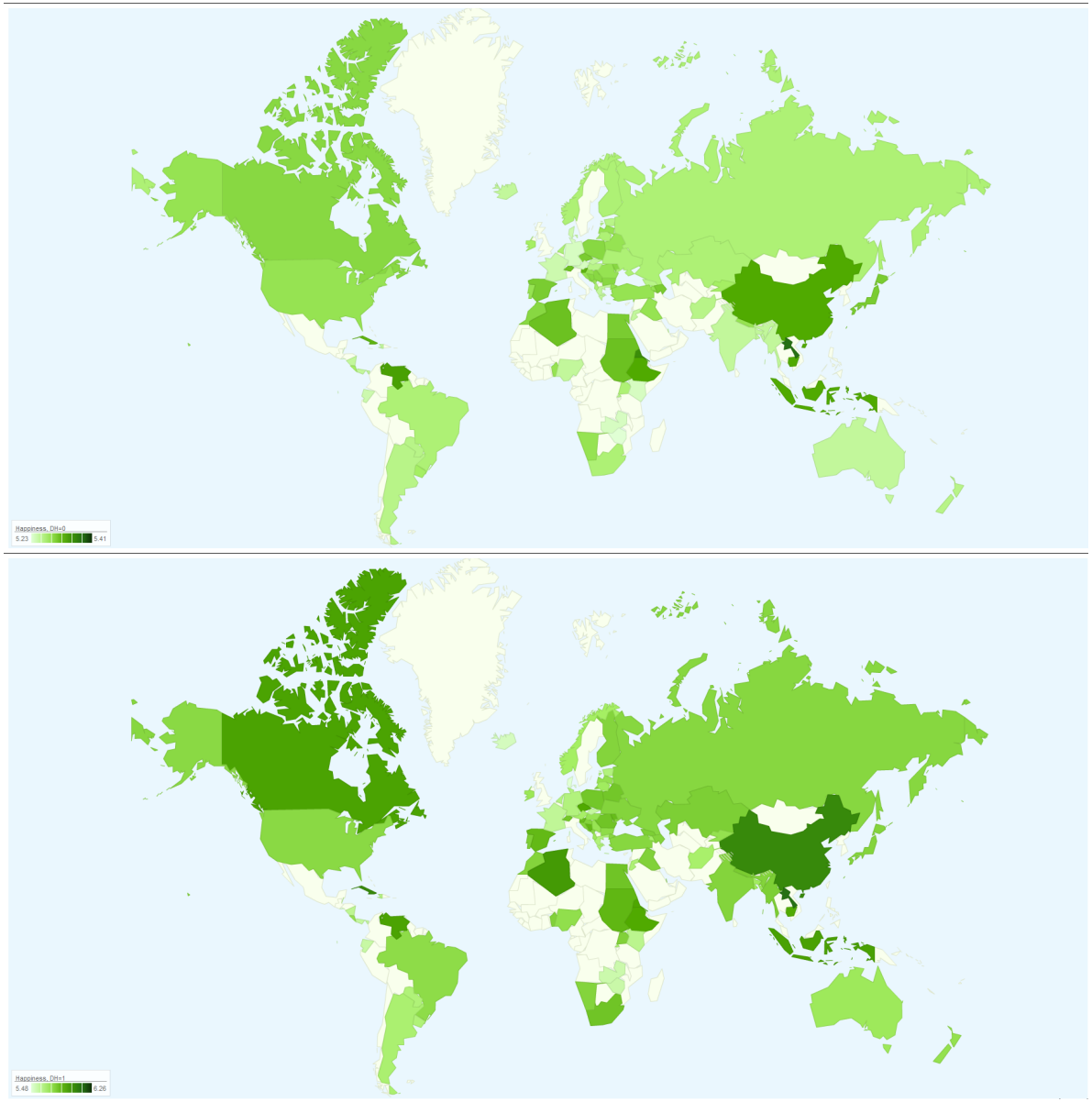

Histogram of Hapiness : $DH = 3, bins = 80$
Number of texts
Level of happiness
0
5
10
15
20
1
2
3
4
5
6
7
8
9

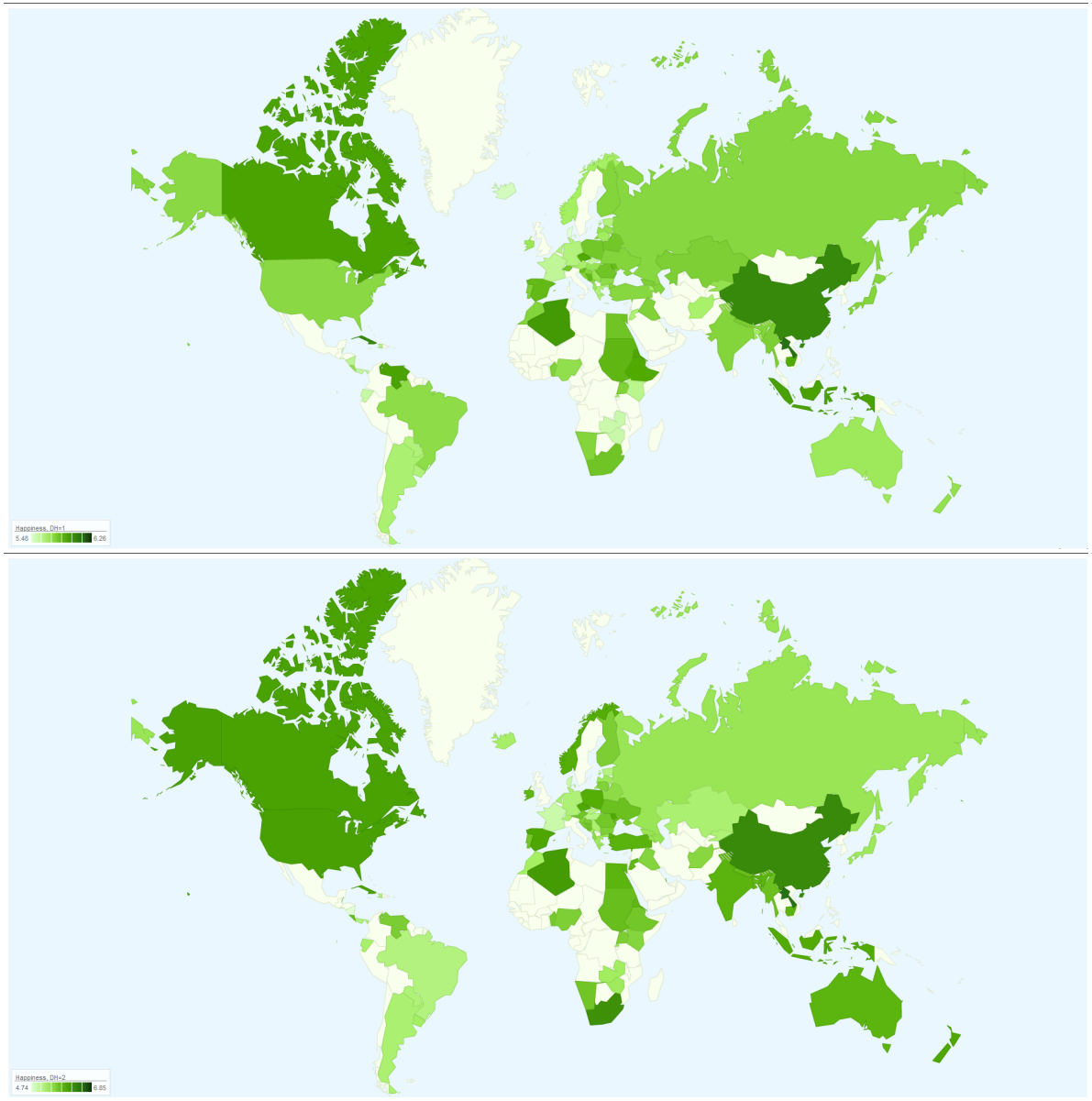

Histogram of Hapiness : $DH = 0, bins = 100$
Number of texts
Level of happiness
0
5
10
15
20
5.25
5.30
5.35
5.40

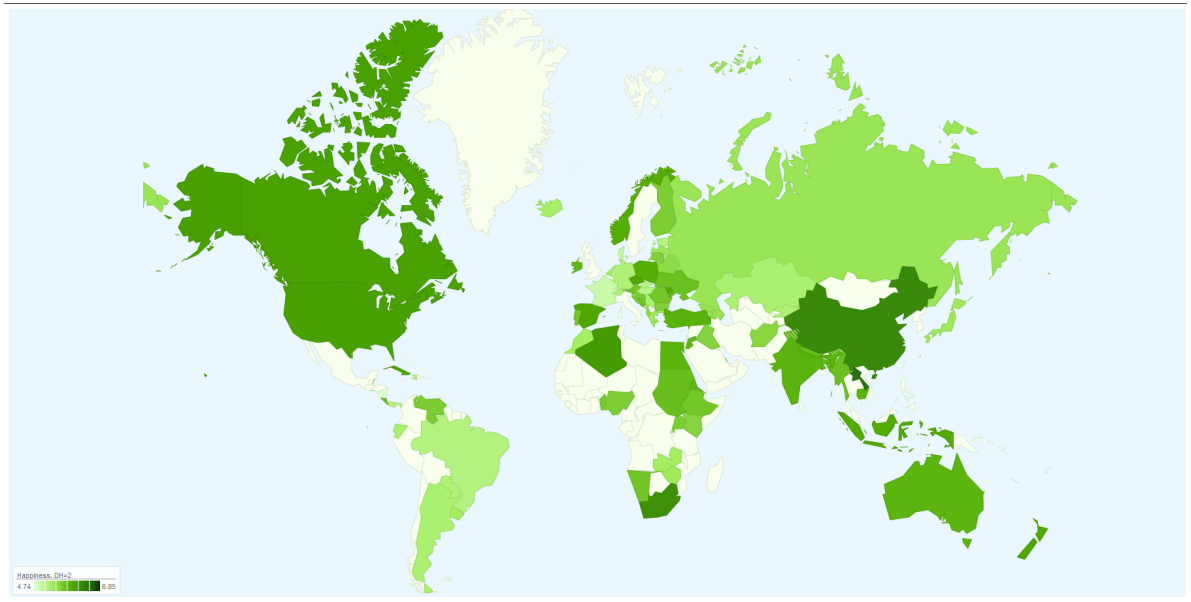

Histogram of Hapiness : $DH = 1, bins = 100$
Number of texts
Level of happiness
0
5
10
15
20
5.6
5.8
6.0
6.2

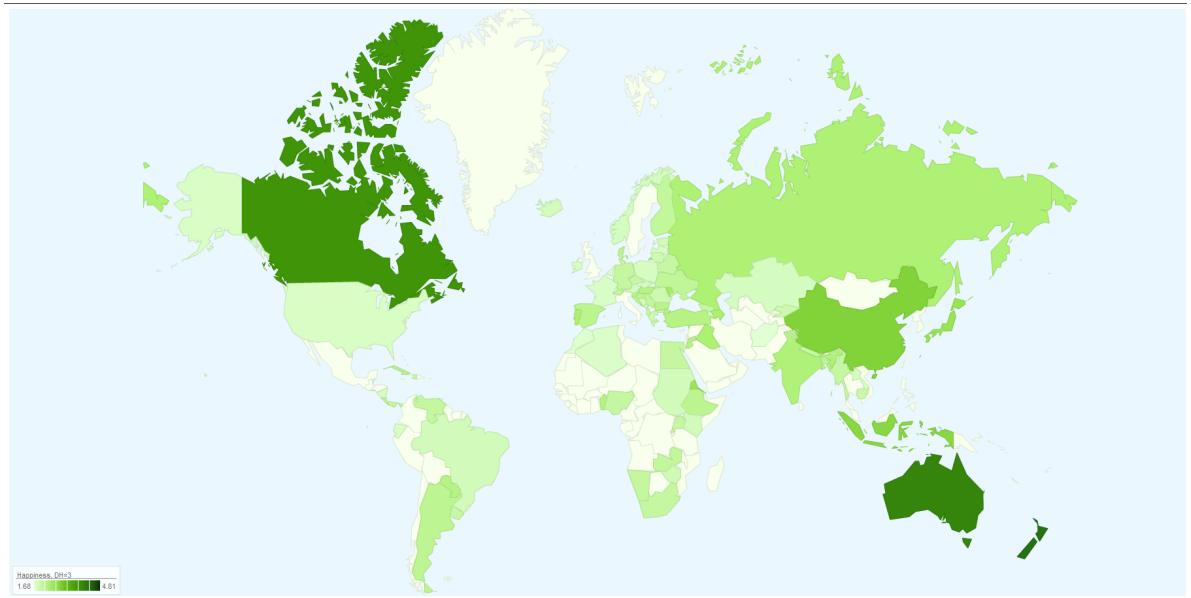

Histogram of Hapiness : $DH = 2, bins = 100$
Number of texts
Level of happiness
0
5
10
15
20
4
5
6
7


Histogram of Hapiness : $DH = 3, bins = 100$
Number of texts
Level of happiness
0
5
10
15
20
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0


Happiness, DH=0
5.23
5.41
Happiness, DH=1
5.48
6.26


Happiness, DH=1
5.48
8.26
Happiness, DH=2
4.74
8.85


Happiness, DH=2
4.74
6.85


Happiness, DH=3
1.68
4.81

Pearson Correlation

Factor	ΔH=0	ΔH=1	ΔH=2	ΔH=3
Limited Government Powers	-0.13	-0.15	0.09	0.24
Absence of Corruption	-0.10	-0.15	0.01	0.26
Order and Security	0.06	0.01	0.03	0.20
Fundamental Rights	-0.11	-0.16	-0.03	0.16
Open Government	-0.11	-0.16	0.04	0.30
Regulatory Enforcement	-0.19	-0.21	-0.04	0.27
Civil Justice	-0.18	-0.19	-0.03	0.20
Criminal Justice	-0.07	-0.02	0.1	0.20

What it means:

Correlation	Negative	Positive
None	-0.09 to 0.0	0.0 to 0.09
Small	-0.3 to -0.1	0.1 to 0.3
Medium	-0.5 to -0.3	0.3 to 0.5
Strong	-1.0 to -0.5	0.5 to 1.0