VIGESIMOSEGUNDO INFORME ESTADO DE LA NACIÓN EN DESARROLLO HUMANO SOSTENIBLE (2015)

Informe Final

Redes Conceptuales y descontento ciudadano

Andrés Segura
Adriana Céspedes

Agosto, 2016

Índice

1. Introducción

El Programa Estado de la Nación (PEN) bajo el marco del Capítulo Especial del XXII Informe se propone estudiar a profundidad el origen del descontento ciudadano con la democracia costarricense. Con el fin de analizar este tema desde una perspectiva automatizada y sistemática se buscan nuevas aproximaciones de estudio que brinden una perspectiva innovadora de la opinión de los ciudadanos sobre el rumbo del país.

El presente trabajo presenta una visualización diferente y novedosa de la opinión pública y el descontento al aplicar técnicas computacionales de recuperación de información y de análisis de redes sociales a las entrevistas efectuadas por el PEN durante el año 2015 a ciudadanos de diferentes grupos demográficos y zonas del país. Con estas técnicas se busca encontrar contra quienes está dirigido el descontento ciudadano.

Se espera que las visualizaciones, métricas y datos obtenidos en este proceso se conviertan en un insumo más que coadyuve a la investigación desarrollada por el PEN para análisis de descontento. Esto debido a que aplicarán su conocimiento experto sobre los productos entregados y con este se dará valor y significado a los productos más allá del quehacer computacional.

En las siguientes secciones del documento se detallarán las nociones básicas y los procesos metodológicos que se siguieron en este trabajo. La sección 2 consiste de un breve marco teórico que aclara los conceptos teóricos requeridos para comprender las bases computacionales aplicadas al análisis de texto. La metodología empleada durante el proceso se encuentra especificada en la sección 3 y finalmente en la sección 4 se muestran los resultados obtenidos a la fecha de presentación de este taller.

2. Marco Teórico

2.1 Recuperación de Información

La Recuperación de Información (RI) es un campo de las ciencias de la computación que se encarga de satisfacer necesidades de información para el usuario mediante la implementación de algoritmos automatizados de extracción de contenido relevante dentro de una colección de documentos o base de datos (Jansen & Rieh, 2010).

Según Baeza-Yates & Ribeiro-Neto (1999), aunque existen diversos modelos de RI, distinguidos unos de otros básicamente por los fundamentos matemáticos que los sustentan (vectoriales, bayesianos, booleanos, entre otros), en general el procedimiento seguido para recuperar un conjunto de documentos basados en un conjunto de palabras de consulta es similar en todos ellos, a saber:

- Análisis léxico: Uniforma el texto de la colección de documentos de forma que se obtengan únicamente palabras con caracteres válidos y con un formato consistente, por ejemplo, todo el texto se convierte a minúscula, se eliminan tildes, números y signos de puntuación.

- Eliminación de Stopwords: En este proceso se elimina de los documentos todas aquellas palabras que no tienen carga semántica en el texto, por ejemplo artículos, preposiciones, entre otros.

- Derivación (Stemming): Para efectos de futuras búsquedas se derivan las raíces de los términos con el fin de agrupar palabras bajo una misma serie de caracteres, por ejemplo la raíz "educ" representaría a educación, educativa, educadoras, etc.

- Indización: Se construye un índice ordenado de los términos presentes en la colección para facilitar las búsquedas.

- Ranking: Se crea un mecanismo que permite posicionar y recuperar los documentos según su relevancia en relación con términos de búsqueda específicos. Dicha relevancia es dada preponderantemente por la frecuencia de los términos en el texto.

En la sección metodológica posterior se mostrará cómo los pasos de RI mencionados se implementan a las entrevistas efectuadas por el Programa del Estado de la Nación para analizar el descontento ciudadano.

2.2 Análisis de Redes Sociales

El Análisis de Redes Sociales (ARS) es en la actualidad una rama importante dentro de las ciencias sociales, la cual ha contribuido con un conjunto de teorías, modelos, metodologías y aplicaciones propias, fundamentados en la comprensión de los grupos sociales como un tejido de relaciones y procesos entre entidades, es decir, como redes sociales (Wasserman & Faust, 1994). El propósito del ARS es brindar un conjunto de métricas que faciliten la descripción y comprensión del comportamiento de las redes sociales.

Dicho análisis asume que las redes sociales poseen una estructura, que se hace evidente en los patrones regulares de interacción entre las entidades concretas (personas, grupos pequeños, organizaciones, entre otros) que participan en ellas (Knoke & Yang, 2008)

Una aplicación reciente y novedosa en esta área se da en el contexto del análisis del flujo conceptual, donde a partir de métodos de ARS es posible construir una red de conceptos con el fin de mostrar contenido latente o emergente que nos es visible mediante técnicas tradicionales de análisis del discurso (Diesner & Carley, 2004).

Cabe aclarar que la red de conceptos se construye mediante algoritmos que sistemáticamente generan relaciones entre las palabras presentes en el texto, asumiendo que, en efecto, el flujo conceptual puede ser derivado de esta manera. El proceso debe ser supervisado mediante criterio experto, ya que la escogencia de los algoritmos de visualización de la red y las métricas de análisis a utilizar dependen del contexto del texto a estudiar (Leydesdorff & Welbers, 2011).

Es importante mencionar que el PEN no es ajeno al uso de ARS en sus investigaciones, ya que recientemente mostró su análisis del estado de la colaboración de las comunidades científicas costarricense a partir de métodos de ARS aplicados a las citas de artículos científicos del país almacenados en las bases de datos de Scopus (PEN, 2014).

La sección de metodología siguiente aclara cómo se utiliza ARS en el contexto específico del análisis del descontento ciudadano.

2. Metodología

En esta sección se detallan cada uno de los procesos involucrados para llegar a los resultados obtenidos. Estos pasos se agrupan en dos fases principales de procesamiento: la primera Recuperación de la Información (RI) y la segunda Análisis de Redes Sociales (ARS).

Como insumos se consideraron 20 encuestas efectuadas durante el 2015 a ciudadanos costarricenses; hombres y mujeres de diferentes edades y zonas del país. Estas encuestas fueron provistas por el Programa del Estado de la Nación y se resumen en la siguiente tabla.

Cuadro 1
Lista de personas entrevistadas y sus características

1. Mujer adulta zona rural. Atenas, Atenas
2. Mujer adulta zona rural. Barbacoas, Puriscal
3. Mujer adulta zona rural. Horquetas, Sarapiquí
4. Mujer adulta zona rural. Tárcoles, Garabito
5. Hombre adulto zona rural. Puraba, Santa Bárbara
6. Hombre joven zona rural. Concepción San Rafael Heredia
7. Hombre joven zona rural. Distrito San Ramón, San Ramón
8. Hombre zona rural. Guácima, Central Alajuela
9. Mujer adulta zona urbana. Carmen, Cantón Central, Cartago

10. Mujer adulta zona urbana. Curridabat, Curridabat
11. Hombre adulto zona urbana. Mercedes Central, Heredia
12. Hombre adulto zona urbana. San Josecito San Rafael Heredia
13. Hombre joven zona urbana. Tres Ríos La Unión
14. Hombre joven zona urbana. Tres Ríos La Unión
15. Mujer joven zona urbana. Distrito Central, Cantón Central
16. Mujer joven zona urbana. Mercedes, Central Heredia
17. Mujer joven zona urbana. Gravilias Desamparados
18. Hombre adulto zona rural.
19. Hombre adulto zona urbana.
20. Mujer joven.

3.1 Fase de Recuperación de Información

El primer paso consiste de un analizador léxico, que permitió generar para cada documento (entrevistas) una lista de términos a considerar para el procesamiento posterior. Esta lista contiene únicamente texto en minúscula, no incluye tildes, valores numéricos y signos de puntuación, igualmente los artículos y preposiciones fueron descartados. El software encargado de realizar este proceso fue desarrollado en su totalidad por el equipo de trabajo del Laboratorio de Investigación e Innovación Tecnológica de la UNED (LIIT)

Posteriormente la lista es depurada mediante ayuda del PEN para eliminar todas aquellas palabras que no poseen carga semántica relevante para el análisis deseado. El conjunto de estas palabras se denomina “stopwords”, o palabras vacías del español. Dentro del conjunto de palabras suprimidas se pueden mencionar los verbos ser y estar; verbos para aseverar criterio personal como “creo”, “pienso” y “muletillas” propias de entrevistas no editadas como: “Por ejemplo”, “Tal vez”, “Por dicha”, “Verdad”, “Mmm”, “Ajá” entre otras. La lista final de palabras suprimidas para esta fase se proveerá con la entrega final.

Seguidamente se procedió a realizar la derivación de los términos, de manera que aquellos semánticamente similares queden agrupados en un único representante. Para esta tarea se implementó el algoritmo de stemming de Martin F. Porter (1980), el cual es uno de los algoritmos más usados y conocidos para derivación de términos.

Al finalizar el punto anterior se cuenta finalmente con una lista de términos representativa de las entrevistas de descontento. Es posible obtener entonces una lista de frecuencias de los términos con menciones de un documento. Además, se obtiene el insumo para la siguiente fase del proceso. La siguiente fase se especifica en la sección 3.2.

3.2 Fase de Análisis de Redes Sociales

Con el fin de poder extraer una red conceptual de los términos obtenidos en la fase previa se siguió un enfoque similar al de Paranyushkin (2011), donde la relación entre los términos se construye según su proximidad en el texto. Así, se debe recorrer todos los términos y a cada uno de ellos se le asignará una relación con al menos dos términos próximos. La proximidad, es decir, la cantidad de palabras a recorrer antes de asignar una relación, estará determinada por el tamaño del salto o distancia deseada entre las palabras según criterio experto.

En el caso particular de esta investigación, los investigadores del PEN decidieron que el salto o distancia a utilizar para generar sistemáticamente las relaciones entre los términos fuera de 3, ya que mostró una mayor expresividad e interpretatividad del objeto de interés durante las primeras pruebas. Esto implica que para cada término de la lista (los cuales pueden repetirse) se establece una relación con los dos términos que se encuentren a 3 palabras de distancia. Si una relación entre términos se repite, lo que sucede es que se le aumenta el peso a esta, como un indicador de fuerza de la relación.

Una modificación nueva a la aproximación de Paranyushkin (2011) fue el uso de un criterio de filtro basado en descriptores provistos por el PEN, y cuya ocurrencia en el texto se conoce a priori (ver Anexo 1). La idea es que los filtros están conceptualizados según categorías de análisis de interés para el informe que esta organización desea generar. Los descriptores fueron provistos por el PEN definiendo palabras “clave” para categorías la creación de nodos y conformación de relaciones.

Consecuentemente, la metodología se modificó para que la red conceptual fuera conformada entonces por los términos asociados a los descriptores brindados aplicando el algoritmo de construcción de relaciones a dos niveles de profundidad, es decir, para cada ocurrencia del descriptor, se crean relaciones con sus correspondientes términos próximos y recursivamente se realiza el procedimiento para dichos dos términos próximos.

Cabe aclarar que este procedimiento fue igualmente programado por el equipo de trabajo y produce como resultado dos archivos, uno que representa los nodos de la red (los términos en este caso) y otro que describe las relaciones entre ellos. El propósito de estos dos archivos es poder suministrar al programa Gephi (Gephi.org, 2016) los datos que faciliten la visualización y el análisis de la red mediante métodos ya validados científicamente, ya que todos los algoritmos utilizados por esta plataforma son referenciados mediante publicaciones científicas indexadas.

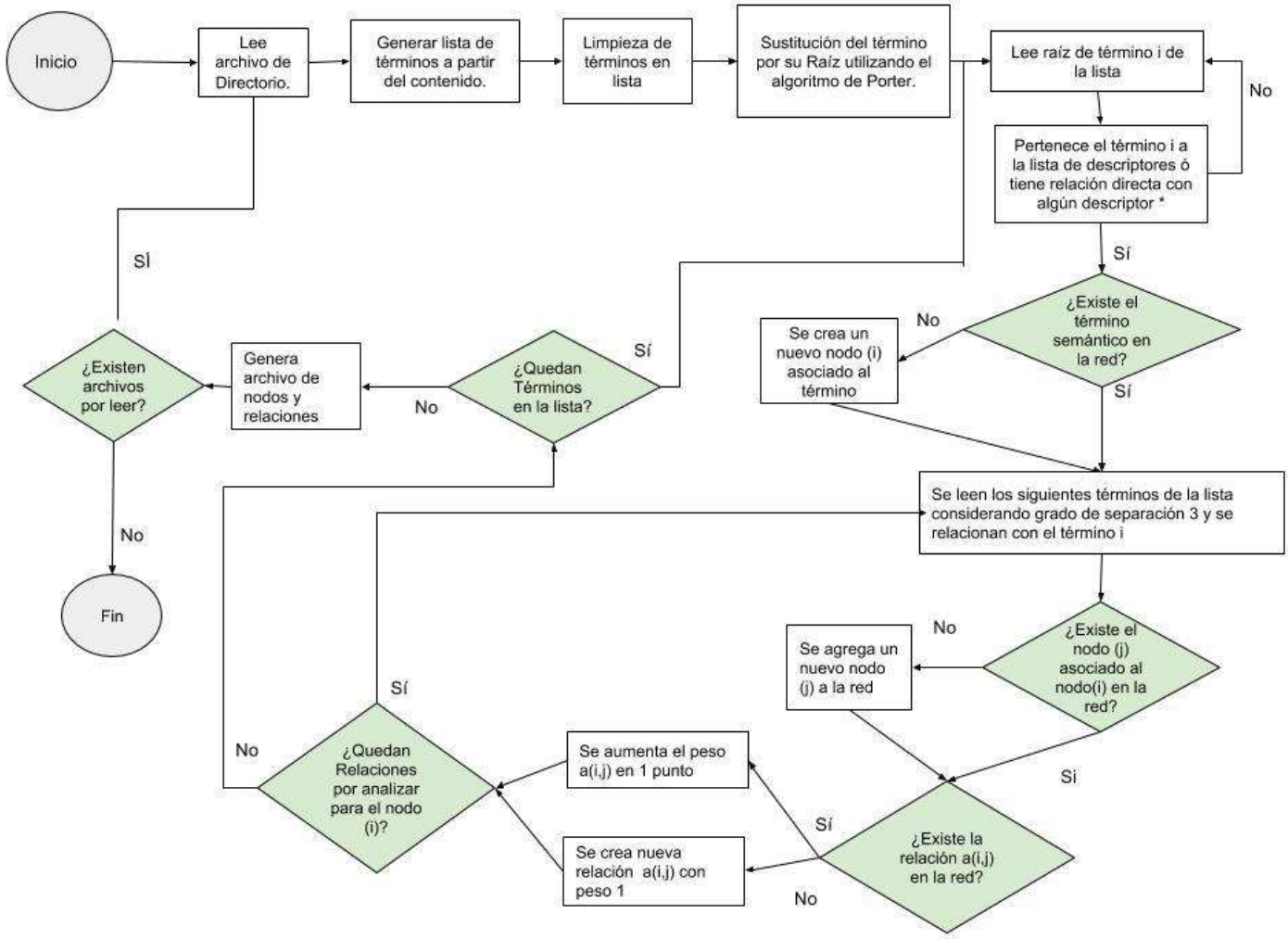
Mediante Gephi se determinó la relevancia de cada nodo de la red aplicando el algoritmo PageRank (Page et. al, 1999), el cual hace un balance entre la cantidad de relaciones que posee el nodo y el peso de cada una de ellas, de manera que se obtiene un indicador de importancia relativa del nodo en la red. Finalmente para efectos de visualización de la red se utilizó el algoritmo Fruchterman & Reinhold (1999) el cual organiza los nodos en función de su relevancia y toma en cuenta de igual forma la cantidad de enlaces con las que el nodo se relaciona y el peso asociado a ellos.

La visualización para entrevistas unificadas se efectuó bajo los siguientes criterios:

a) Visualización del 5% de los nodos con valor de PageRank más alto (nodos más relevantes)
b) Los colores de los nodos se representan con una escala de Amarillo-Rojo, dónde amarillo corresponde a los Page Ranks más bajos y rojo a los valores más altos.
c) El tamaño de los nodos se grafica de acuerdo al valor del PageRank con una proporción de tamaño de escala 10 – 120. (El nodo con menor Page Rank graficado con tamaño 10 y el nodo con mayor tamaño graficado de tamaño 120)

La figura 1 muestra el diagrama de flujo de la metodología incluyendo ambas fases.

Figura 1
Diagrama de flujo de la metodología empleada.



Fuente: Elaboración propia.

En la sección 4, se presentarán las visualizaciones y resultados obtenidos del análisis de las entrevistas.

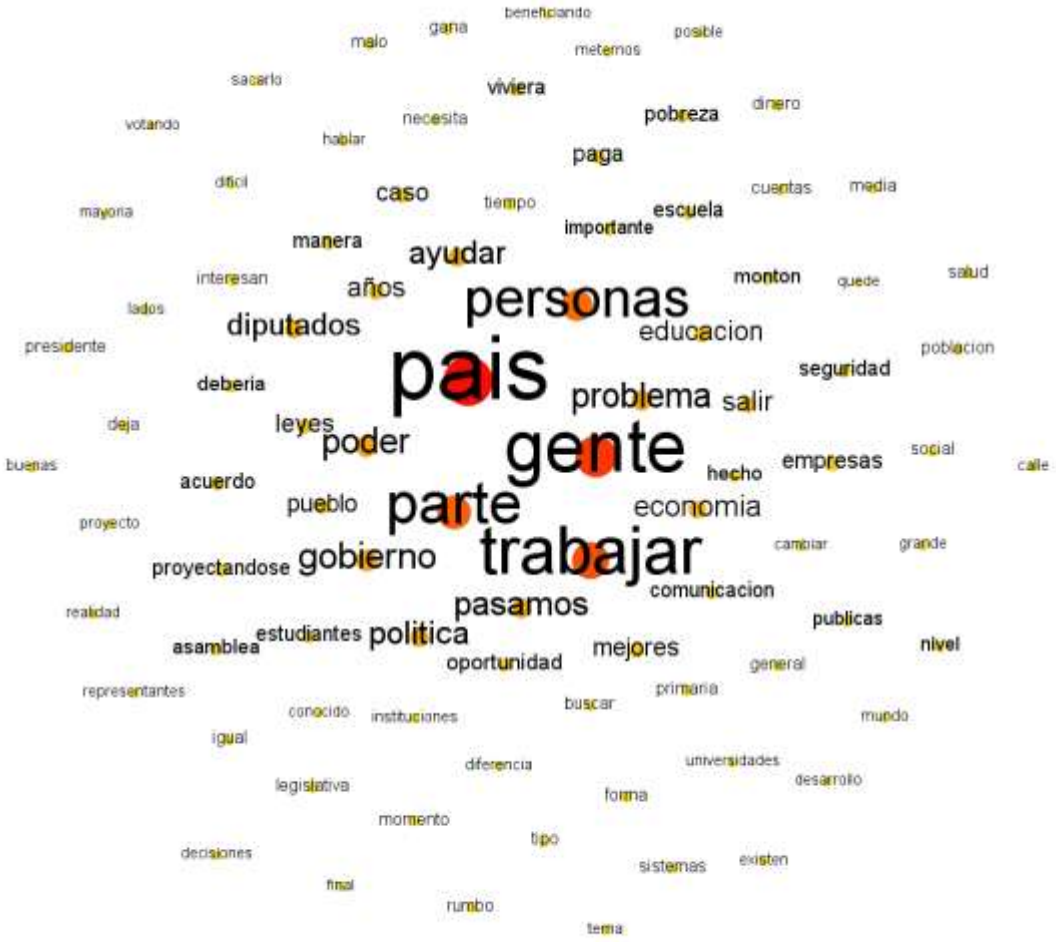
4. Resultados

Cuadro 2
Resumen de Nodos en la Red

Entrada: Entrevistas Unificadas en un sólo archivo de insumo.	
Total de Nodos en la Red:	**4446**
Nodos Considerados para visualización (5%)	**89**

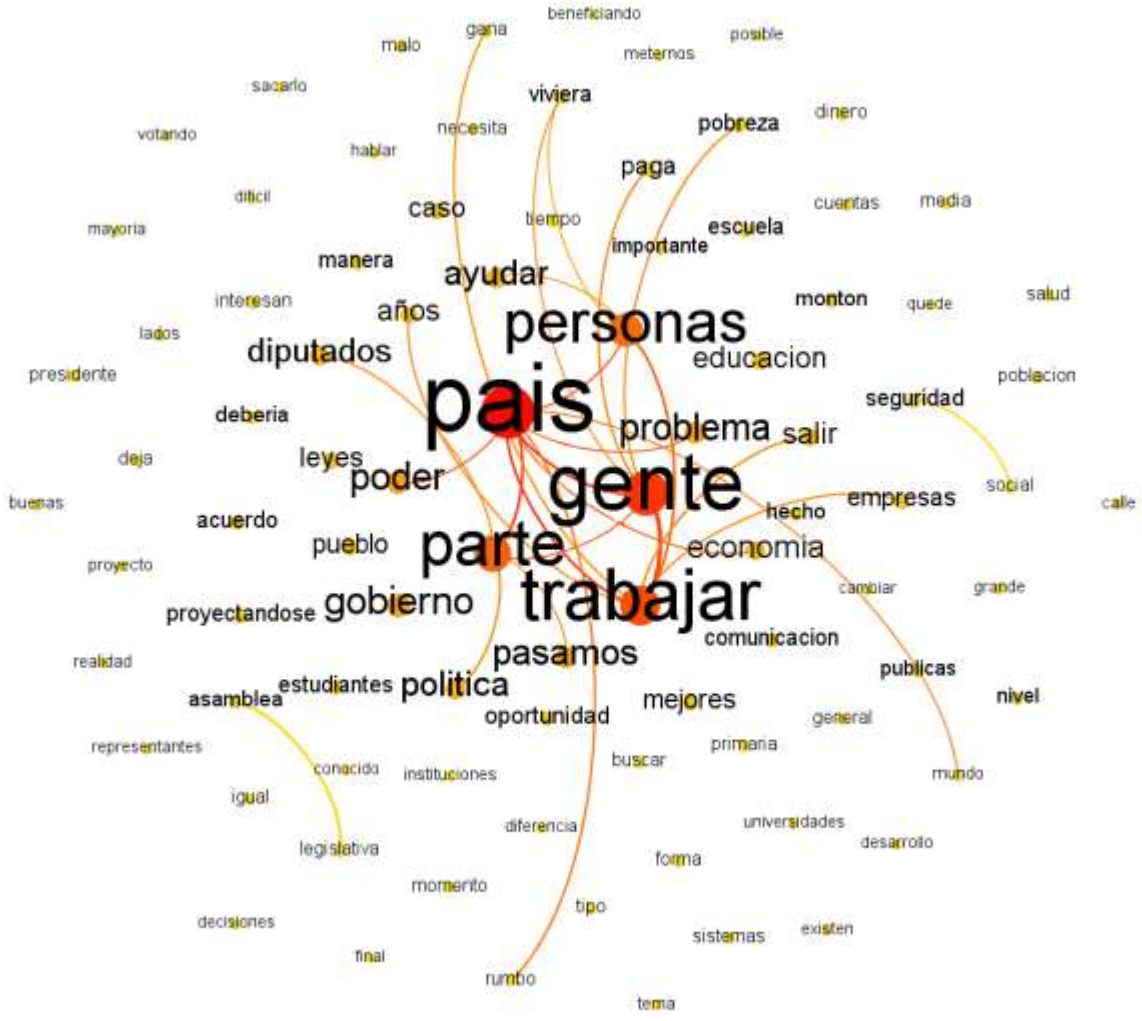
Fuente: Elaboración propia

Figura 2
Red conceptual de las Entrevistas de Descontento Unificadas Sin Aristas.



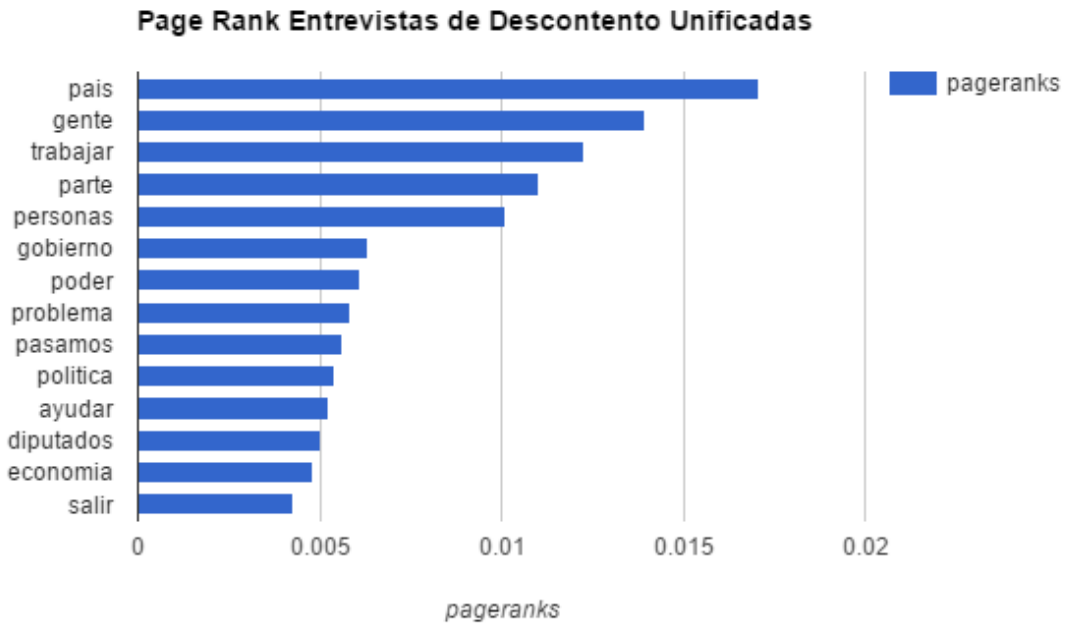
Fuente: Elaboración propia

Figura 3
Red de flujo conceptual para las entrevistas de descontento unificadas con aristas de Peso mayor a 15 en un rango de 15-39. El peso identifica la cantidad de menciones



Fuente: Elaboración propia

Gráfico 1
Page Rank para las entrevistas de descontento unificadas, con los 15 términos con valor más alto.



Fuente: Elaboración propia

Cuadro 3
Page Rank para las entrevistas de descontento unificadas. Corresponde a los 89 Términos más preponderantes.

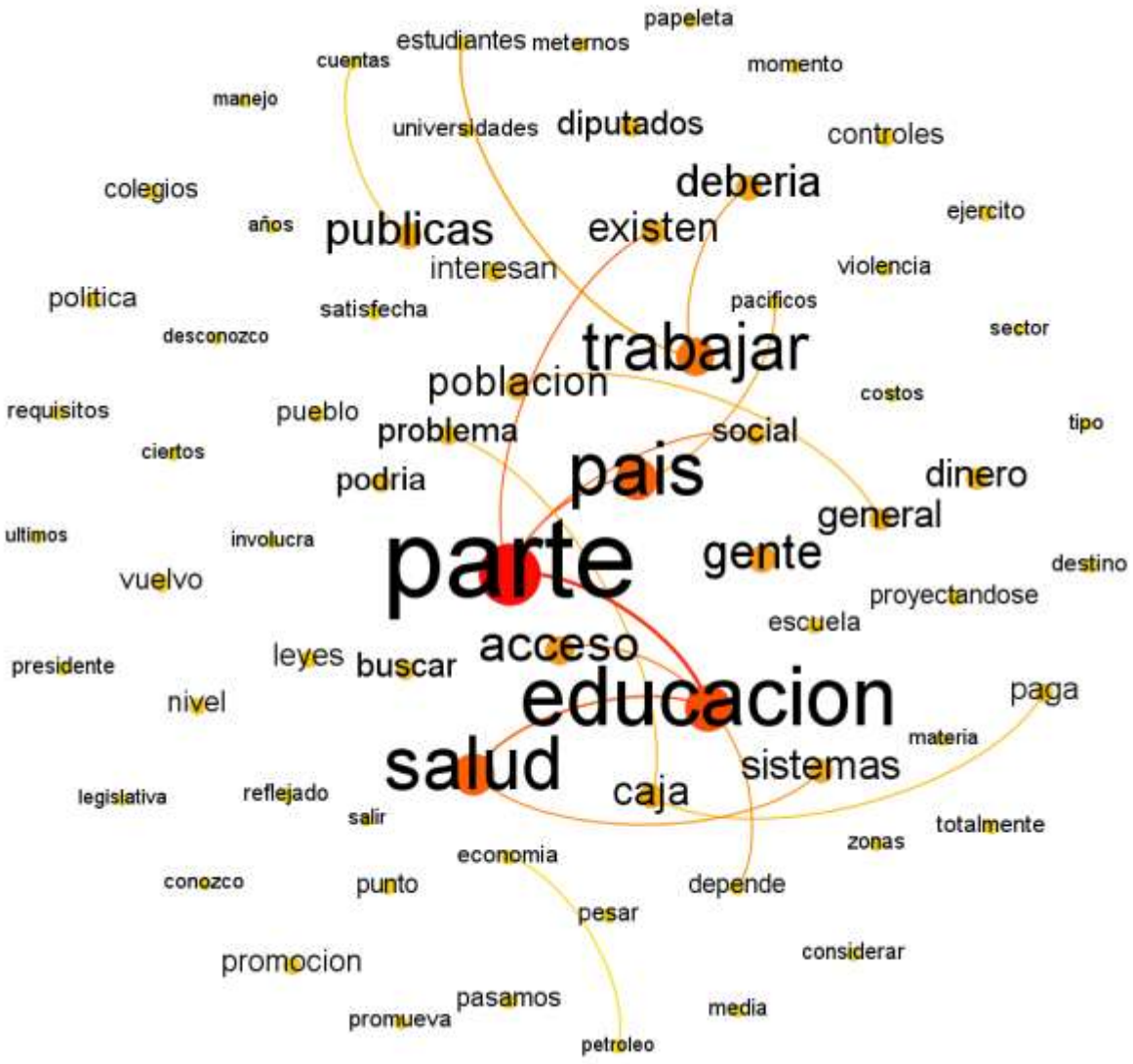
Lista de términos visualizados con su valor de Page Rank	
Concepto	**PageRanks**
país	0.01709380834
gente	0.01396984432
trabajar	0.0122707142
parte	0.01100215879
personas	0.01009292905
gobierno	0.006297362589
poder	0.006124793463
problema	0.005835685643
pasamos	0.00563325199
política	0.005399838053

ayudar	0.005232391944
diputados	0.005044101815
economía	0.00481073836
salir	0.004285362735
educación	0.004190271407
años	0.004175216001
leyes	0.003824959991
caso	0.003661023154
pueblo	0.003626669869
mejores	0.003607349285
empresas	0.003378911019
paga	0.003334787566
oportunidad	0.003227590767
proyectándose	0.003024806238
acuerdo	0.003019262325
estudiantes	0.003008020686
comunicación	0.003003657034
pobreza	0.002993973893
seguridad	0.002968267774
manera	0.002952705357
viviera	0.0028981211
escuela	0.002888045139
hecho	0.002778326887
montón	0.002679978777
asamblea	0.002633265642
debería	0.002604176093
importante	0.002594384473
publicas	0.002526501401
nivel	0.002423240193
tiempo	0.002353477818
interesan	0.002338258256
dinero	0.002303562417
salud	0.002293195469
gana	0.002283450212
presidente	0.002261177542
legislativa	0.002224606188
necesita	0.002211021013
momento	0.002196589322
malo	0.002187142562
cuentas	0.002178826418

general	0.002098041525
forma	0.002090605776
sistemas	0.002079679264
rumbo	0.002051127472
social	0.002034533415
buscar	0.0020313961
igual	0.002031112151
tipo	0.002017421931
deja	0.001995375848
primaria	0.001988588865
media	0.001985874967
población	0.001964605002
realidad	0.001955140044
calle	0.001949500364
meternos	0.00194944651
lados	0.001901500226
instituciones	0.001893728672
proyecto	0.001893446926
buenas	0.001886638282
diferencia	0.001868366231
hablar	0.001841285343
tema	0.001828486511
mundo	0.001805373507
grande	0.001785157295
posible	0.001755934799
existen	0.001738307574
sacarlo	0.001735532438
desarrollo	0.001688226412
universidades	0.001685156132
quede	0.001670248549
mayoría	0.001668040425
decisiones	0.001659045945
final	0.00160628528
cambiar	0.001606192527
representantes	0.001602290081
beneficiando	0.001598781028
difícil	0.001588819134
votando	0.001586962285
conocido	0.001580583206

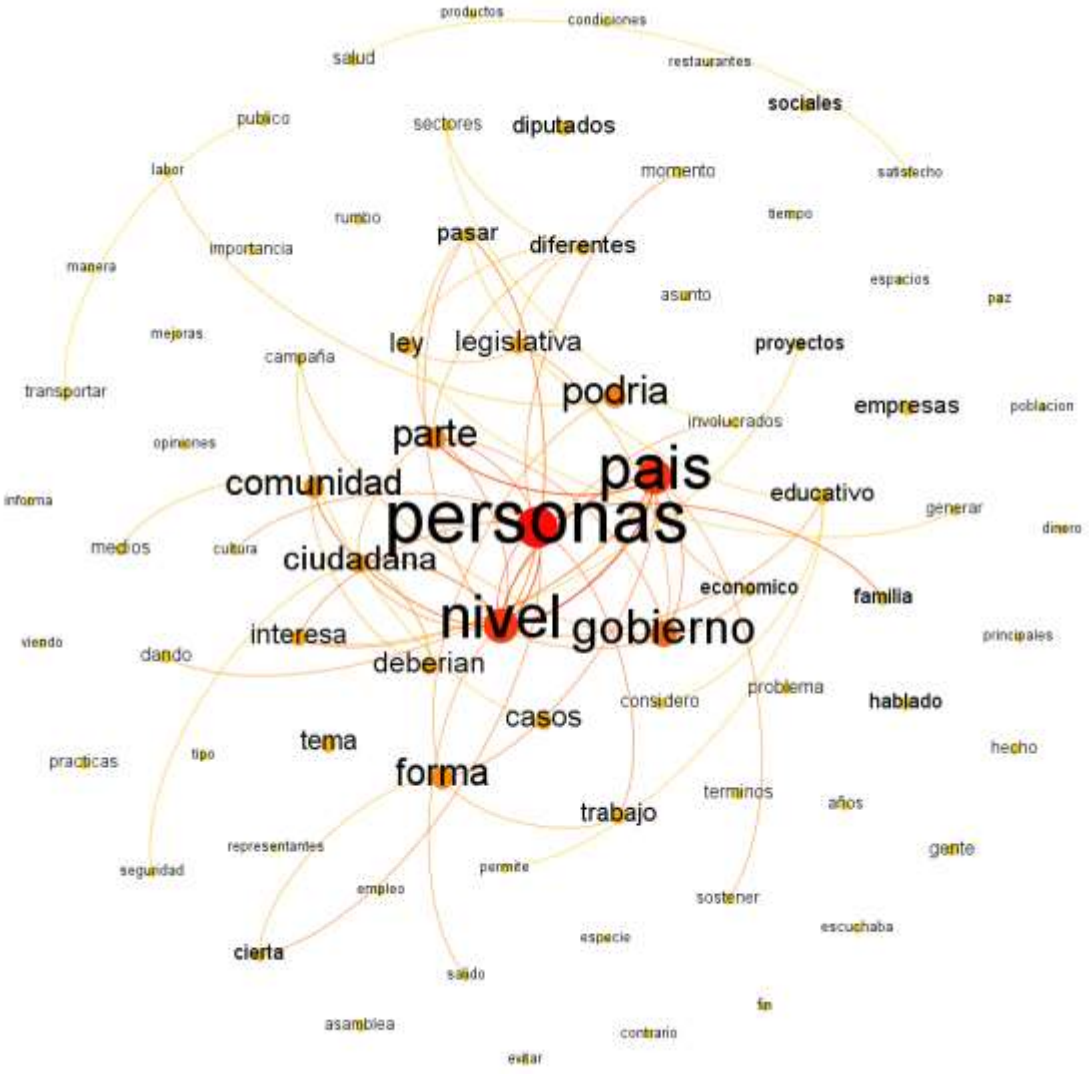
Fuente: Elaboración propia

Figura 4
Ejemplo de Red Individual Mujer Joven. Arista Peso 3-6. Total Nodos 694. Porcentaje visualizado 10%



Fuente: Elaboración propia

Figura 5
Ejemplo de Red Individual Hombre Adulto Urbano. Aristas Peso 2-4.
Total de Nodos: 775. Porcentaje visualizado 10%



Fuente: Elaboración propia

5. Referencias

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval* (Vol. 463). New York: ACM press.

Diesner, J., & Carley, K. M. (2004). Using network text analysis to detect the organizational structure of covert networks. In *Proceedings of the North American Association for Computational Social and Organizational Science (NAACSOS) Conference*.

Fruchterman, T. M. J., & Reingold, E. M. (1991). *Graph Drawing by Force-Directed Placement*. Software: Practice and Experience, 21(11).

Gephi.org (2016) *Gephi: The Open Graph Viz Platform*. https://gephi.org/

Jansen, B. J., & Rieh, S. Y. (2010). The seventeen theoretical constructs of information searching and information retrieval. *Journal of the American Society for Information Science and Technology*, 61(8), 1517-1534.

Knoke, D., & Yang, S. (2008). *Social network analysis* (Vol. 154). London, Sage.

Leydesdorff, L., & Welbers, K. (2011). The semantic mapping of words and co-words in contexts. *Journal of Informetrics*, 5(3), 469-475.

Page, L., Brin, S., Motwani, R. & Winograd, T. (1999) *The PageRank Citation Ranking: Bringing Order to the Web*. Technical Report. Stanford InfoLab.

Paranyushkin, D. (2011). Identifying the pathways for meaning circulation using text network analysis. *Berlin: Nodus Labs*. Recuperado de: http://noduslabs.com/research/pathways-meaning-circulation-text-network-analysis.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130-137.

Programa Estado de la Nación (PEN) (2012). *Decimonoveno Informe Estado de la Nación*. San José, Costa Rica.

Programa Estado de la Nación (PEN) (2014). *Estado de la Ciencia, la Tecnología y la Innovación*. San José, Costa Rica.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications* (Vol. 8). Cambridge University Press.

Anexo 1

Descriptores descontento definidos por el PEN

1. hac
2. pais
3. pued
4. gent
5. person
6. trabaj
7. part
8. bien
9. cost
10. pod
11. educ
12. mejor
13. gobiern
14. problem
15. ayud
16. diput
17. proyect
18. polit
19. igual
20. mal
21. econom
22. president
23. public
24. social
25. asambl
26. ley
27. puebl
28. empres
29. escuel
30. segur
31. acuerd
32. cambi
33. necesit
34. oportun
35. comun
36. legisl
37. satisfech
38. dificil
39. salud
40. corrupcion

41. rumb
42. hag
43. pong
44. poblacion
45. ciudad
46. represent
47. aprob
48. institu
49. aument
50. democraci
51. decision
52. desarroll
53. costarricens
54. democrat
55. libert
56. pacif
57. violenci
58. solucion
59. insatisfech
60. municipal
61. particip
62. ministeri
63. program
64. infraestructur
65. preocup
66. arregl
67. desempl
68. esfuerz
69. corrupt
70. eleccion
71. desigualdad