Trabajo Práctico Final
Análisis completo de un conjunto de datos

AGUSTÍN MISTA
Universidad Nacional de Rosario
Tópicos de Minería de Datos
Rosario, 28 de Diciembre de 2017

Origen de los datos

Para este trabajo final utilizamos el dataset *Spam Base*[1], orientado a la clasificación de emails en deseados (*Ham*) y no-deseados (*Spam*). El mismo fue creado utilizando una base de datos de emails reales de la compañia Hewlett-Packard en 1998 y estudiado en el artículo Spam![2].

El concepto de Spam es diverso: publicidades de productos/páginas web, esquemas piramidales, cadenas de correo, pornografía, etc. La prevención (mediante filtrado) de esta clase de emails es una tarea que involucra reconocer ciertos patrones en el contenido de los mails entrantes. Debido a ésto, resulta interesante poseer de un conjunto previamente clasificado de correo entrante del que podamos extraer estos patrones con el fin de crear filtros de Spam personalizados, filtrando efectivamente todo el correo no deseado, con el mínimo número posible de falsos positivos.

Este dataset cuenta con 4601 muestras, cada una con 57 features que describen la frecuencia de aparición de ciertas palabras (prefijo `word`) o caracteres (prefijo `char`) clave obtenidos a partir de los emails analizados, junto con las frecuencias de aparición de secuencias de caracterese capitalizados (prefijo `capital`).

A continuación se muestra la lista de features de este dataset, ordenadas según el índice de su respectiva columna en los datos.

1. `word_make`
2. `word_address`
3. `word_all`
4. `word_3d`
5. `word_our`
6. `word_over`
7. `word_remove`
8. `word_internet`
9. `word_order`
10. `word_mail`
11. `word_receive`
12. `word_will`
13. `word_people`
14. `word_report`
15. `word_addresses`
16. `word_free`
17. `word_business`
18. `word_email`
19. `word_you`
20. `word_credit`
21. `word_your`
22. `word_font`
23. `word_000`
24. `word_money`
25. `word_hp`
26. `word_hpl`
27. `word_george`
28. `word_650`
29. `word_lab`
30. `word_labs`
31. `word_telnet`
32. `word_857`
33. `word_data`
34. `word_415`
35. `word_85`
36. `word_technology`
37. `word_1999`
38. `word_parts`
39. `word_pm`
40. `word_direct`
41. `word_cs`
42. `word_meeting`
43. `word_original`
44. `word_project`
45. `word_re`
46. `word_edu`
47. `word_table`
48. `word_conference`
49. `char_;`
50. `char_(`
51. `char_[`
52. `char_!`
53. `char_$`
54. `char_#`
55. `capital_average`
56. `capital_longest`
57. `capital_total`

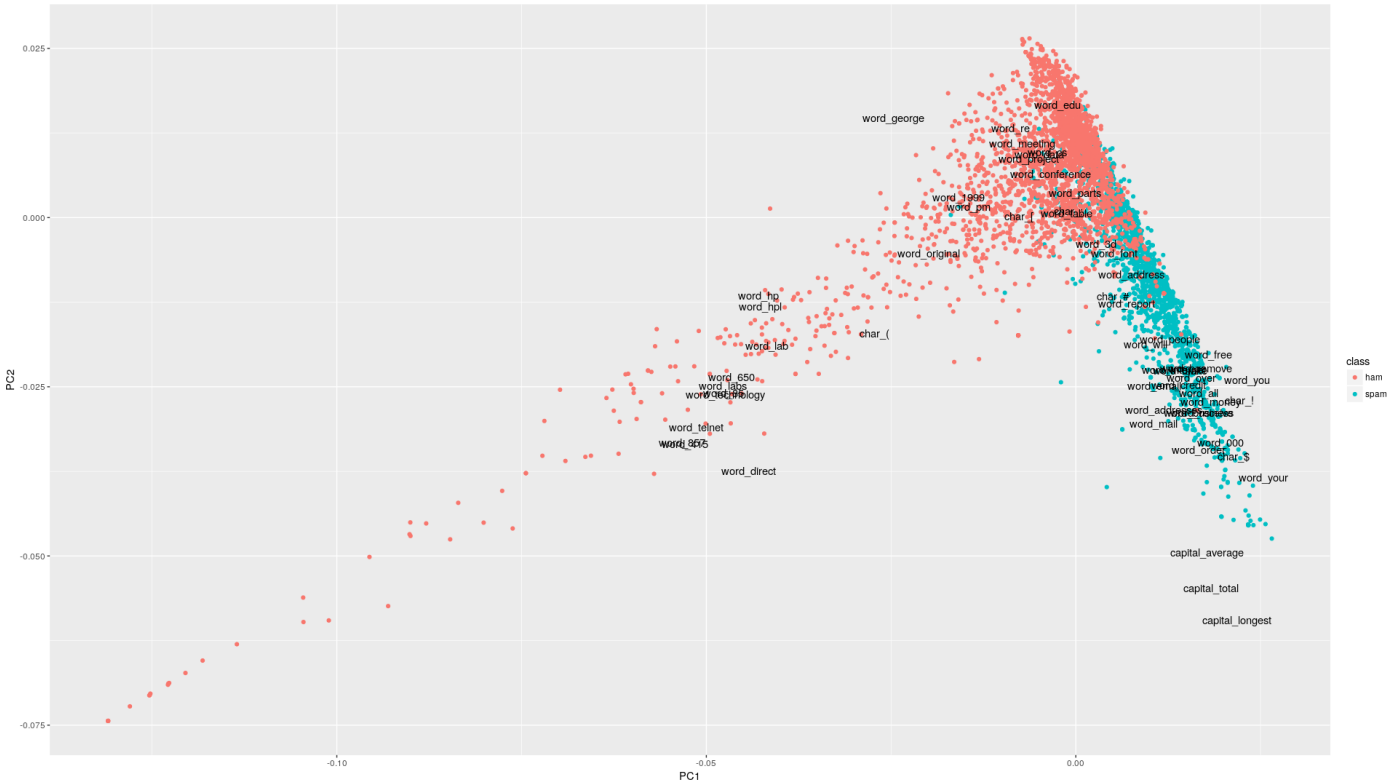
Preprocesamiento En las siguientes secciones se muestran resultados obtenidos a partir tanto del dataset original, como de alguna combinación de preprocesamientos logarítmico (`log()`), escalado (`scale()`) o Principal Component Analysis (`prcomp()`), los cuales se detallan en cada caso particuar.

Visualización de los datos

Ya que este dataset cuenta con un gran número de features, la manera más intuitiva de visualizar los datos en un gráfico de menor dimensionalidad es efectuar una PCA sobre los mismos. A continuación se muestra este análisis sobre nuestros datos, utilizando colores para separar las clases de emails y etiquetas para señalar hacia dónde y con qué relevancia aporta evidencia cada feature.

[1] `https://archive.ics.uci.edu/ml/datasets/spambase`

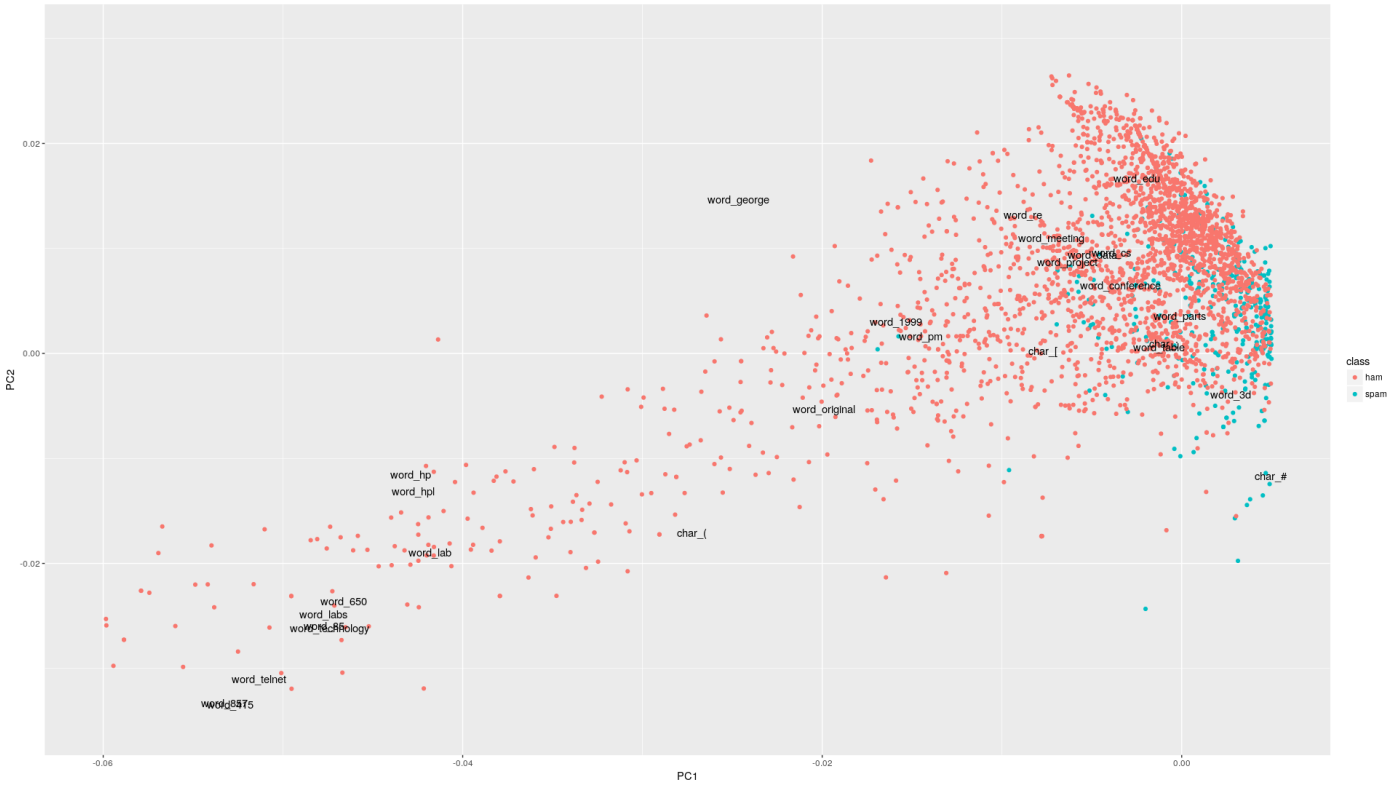
[2] Cranor, Lorrie F., LaMacchia, Brian A. Spam! Communications of the ACM, 41(8):74-83, 1998.



Principal Component Analysis

A primera vista, podemos notar la existencia de dos ejes sobre los cuales se situan las features más relevantes de nuestro dataset, y que se corresponden con aquellos emails deaseados o basura cada caso. Ambos ejes se se unen en el origen de nuestro gráfico, formando una nube densa de muestras donde no existe suficiente evidencia por parte de las features de menor relevancia como para lograr separar cada muestra según su clase de manera efectiva. Suponemos que la mayoría de los errores de clasificación deberían producirse en esa zona.

Si ampliamos el eje correspondiente al correo deseado, podemos observar como la presencia de palabras como “telnet”, “lab”, “hp” o códigos de área telefónica son buenos indicadores de que estamos ante un correo deseado. Resulta evindente que estas palabras están relacionadas directamente con los intereses de la empresa de donde provienen los emails analizados, por lo que debería usarse un dataset de origen más amplio si se desea construir un filtro de Spam efectivo en entornos más diversos.



Si en cambio ampliamos el eje correspondiente al correo no deseado, podemos observar como la presencia de palabras como “money”, “free”, “your”, caracteres como “$”, “!” o largas secuencias de caracteres capitalizados son buenos indicadores de correo basura.

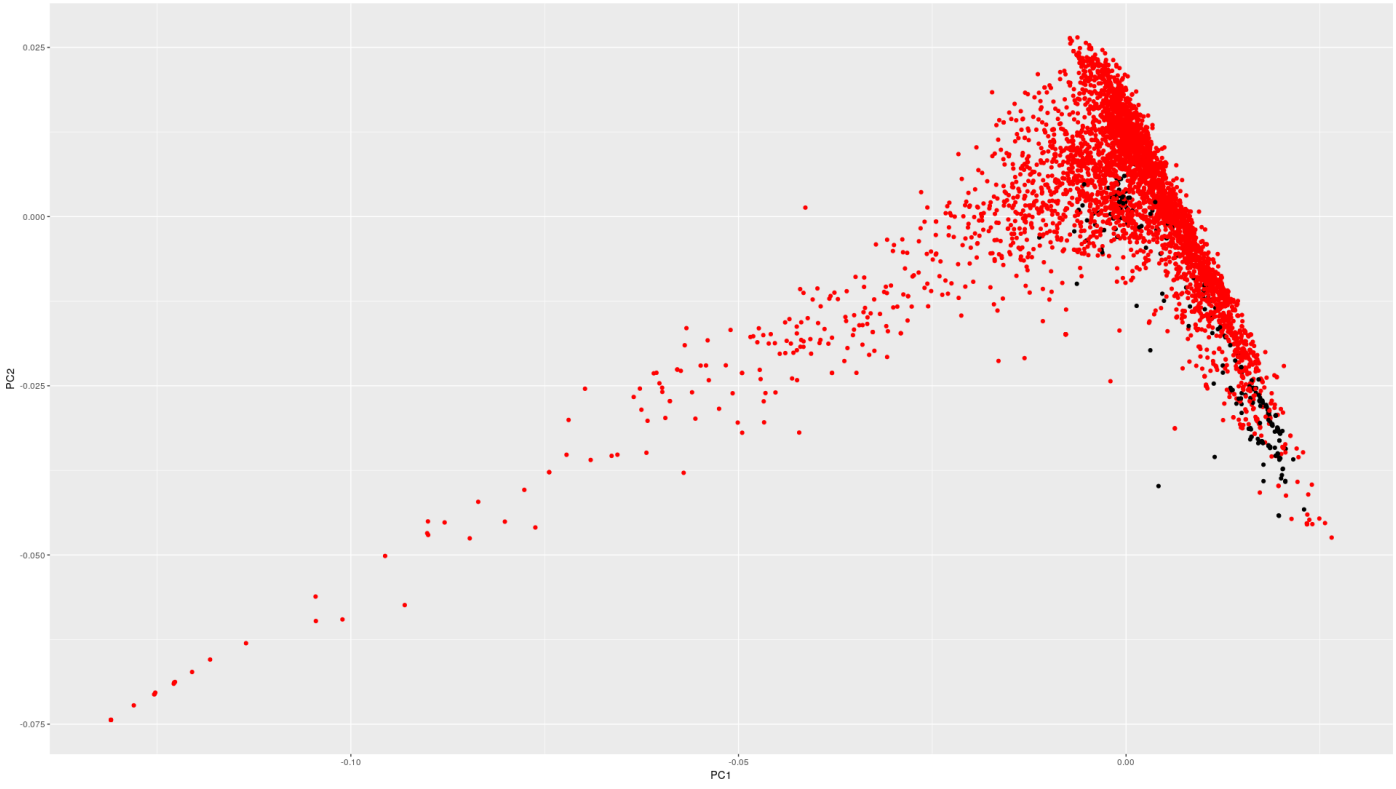
Análisis de features relevantes

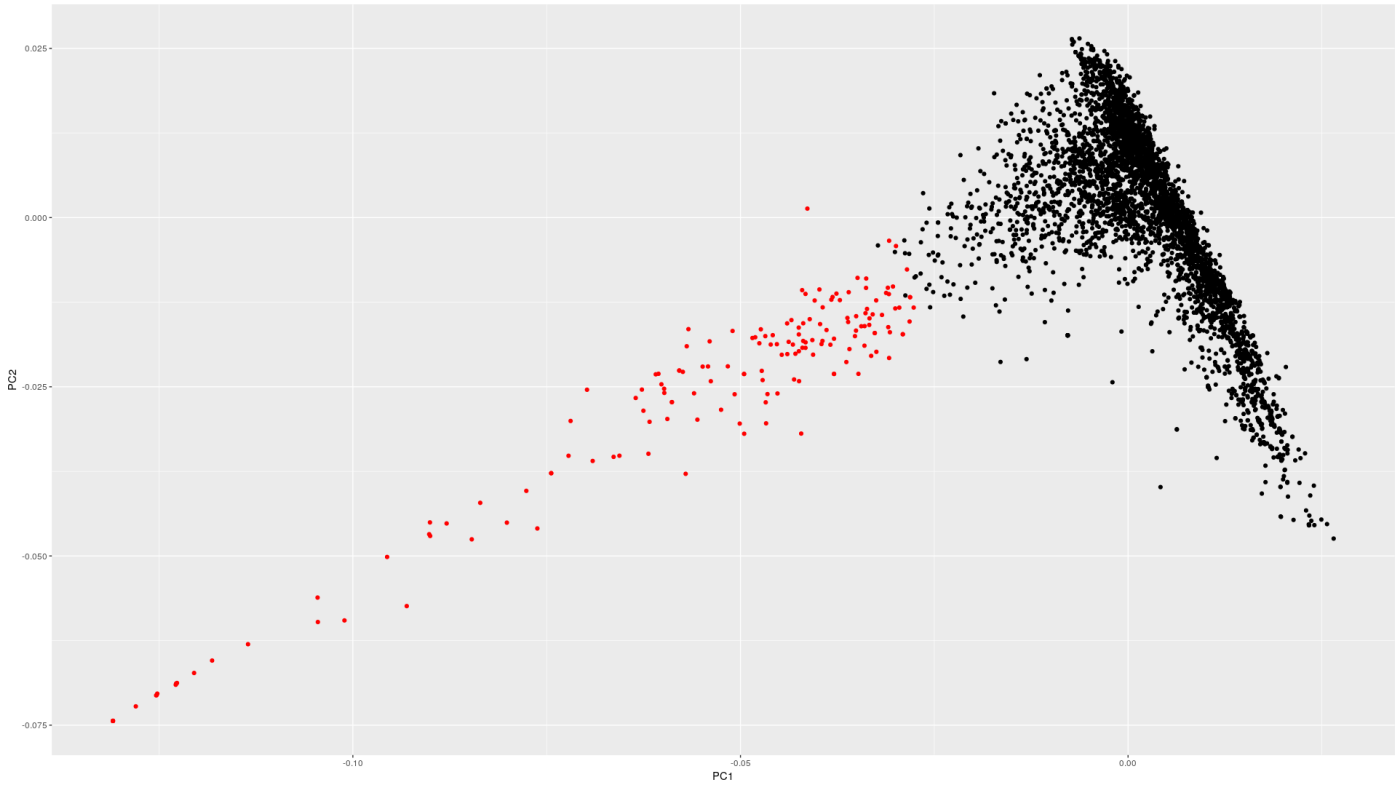
Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Donec hendrerit tempor tellus. Donec pretium posuere tellus. Proin quam nisl, tincidunt et, mattis eget, convallis nec, purus. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Nulla posuere. Donec vitae dolor. Nullam tristique diam non turpis. Cras placerat accumsan nulla. Nullam rutrum. Nam vestibulum accumsan nisl.

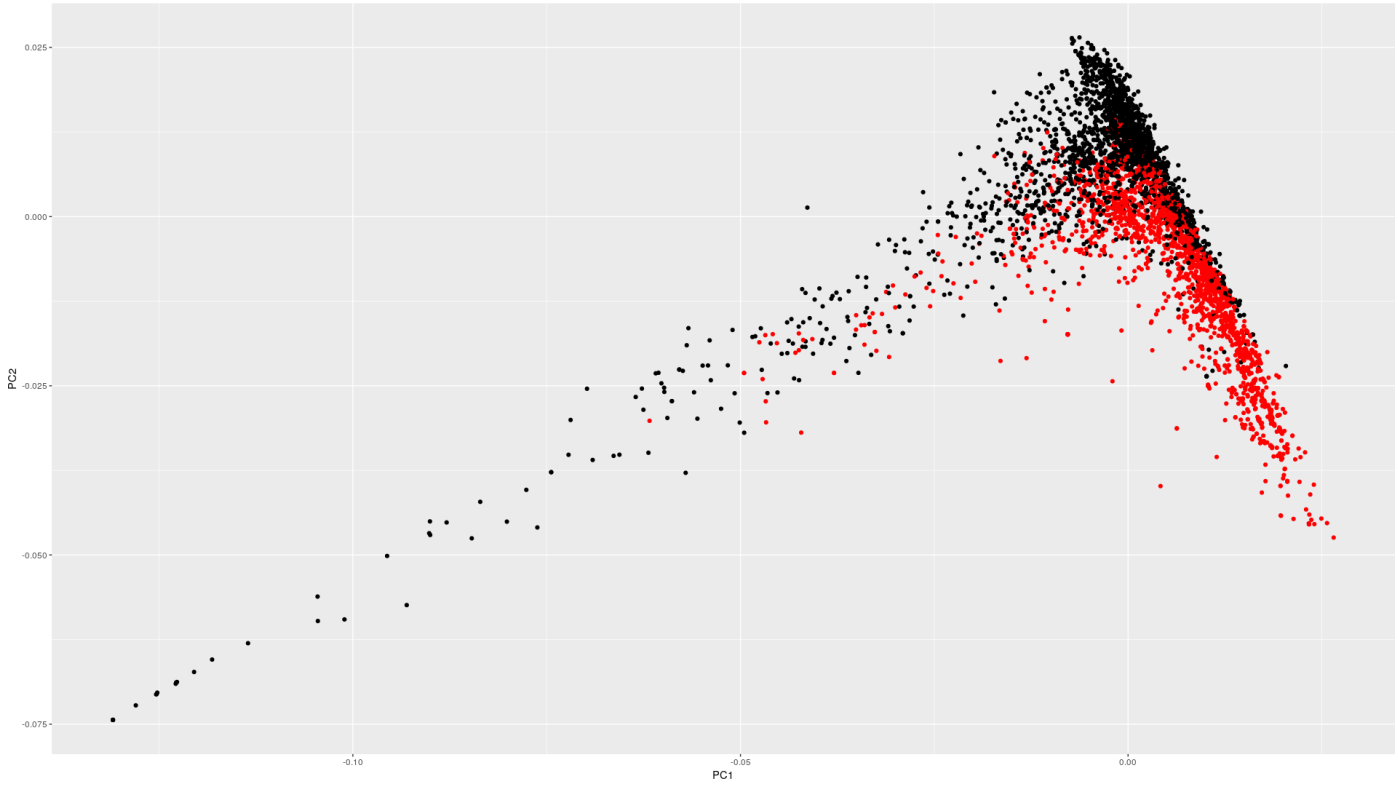
Forward Selection	RF	53	7	52	57	19	50	16	56	25	46
	LDA	53	52	56	7	25	8	5	21	46	42
	SVM	53	7	52	25	57	46	16	21	5	8
Backward Elimination	RF	53	7	52	56	25	5	46	27	19	16
	LDA	52	57	7	25	24	23	16	21	46	8
	SVM	53	7	52	25	57	46	5	27	42	16
Recursive Feature Elimination	RF	7	25	53	52	55	16	46	21	27	5
	SVM	27	41	25	46	26	42	57	53	7	16
Kruskal-Wallis		52	53	7	56	16	21	55	24	57	23

Clustering

Pellentesque dapibus suscipit ligula. Donec posuere augue in quam. Etiam vel tortor sodales tellus ultricies commodo. Suspendisse potenti. Aenean in sem ac leo mollis blandit. Donec neque quam, dignissim in, mollis nec, sagittis eu, wisi. Phasellus lacus. Etiam laoreet quam sed arcu. Phasellus at dui in ligula mollis ultricies. Integer placerat tristique nisl. Praesent augue. Fusce commodo. Vestibulum convallis, lorem a tempus semper, dui dui euismod elit, vitae placerat urna tortor vitae lacus. Nullam libero mauris, consequat quis, varius et, dictum id, arcu. Mauris mollis tincidunt felis. Aliquam feugiat tellus ut neque. Nulla facilisis, risus a rhoncus fermentum, tellus tellus lacinia purus, et dictum nunc justo sit amet elit.







Clasificación

Nullam eu ante vel est convallis dignissim. Fusce suscipit, wisi nec facilisis facilisis, est dui fermentum leo, quis tempor ligula erat quis odio. Nunc porta vulputate tellus. Nunc rutrum turpis sed pede. Sed bibendum. Aliquam posuere. Nunc aliquet, augue nec adipiscing interdum, lacus tellus malesuada massa, quis varius mi purus non odio. Pellentesque condimentum, magna ut suscipit hendrerit, ipsum augue ornare nulla, non

luctus diam neque sit amet urna. Curabitur vulputate vestibulum lorem. Fusce sagittis, libero non molestie mollis, magna orci ultrices dolor, at vulputate neque nulla lacinia eros. Sed id ligula quis est convallis tempor. Curabitur lacinia pulvinar nibh. Nam a sapien.