Text Mining - Trabajo Final

Clasificación de Anotaciones para una Empresa Agropecuaria

Grupo 1: Grijalba, Richard, Ruiz, Szekieta

Sobre la Empresa

Syngenta es una de las compañías líderes del mundo con más de 28.000 colaboradores en más de 90 países.

OBJETIVO

Proporcionarle más seguridad alimentaria de manera sostenible para el medio ambiente a un mundo cada vez más poblado, mediante la creación de un cambio mundial escalonado en la producción agrícola.

En Argentina

En la Argentina, emplea a más de 700 colaboradores e incorpora a más de 2.000 personas para las temporadas de producción.

A pesar de realizar muchos esfuerzos en I+D para semillas y otros productos tecnológicos, Syngenta no tiene experiencia en realizar desarrollos en manejo de información tecnológica.En los últimos años se destinaron inversiones millonarias en productos y equipo para comenzar a desarrollar una nueva área digital para mantener el futuro del negocio que se está viendo afectado por las tendencias mundiales de generar alimentos cada vez en mayor escala a una utilización de los recursos necesarios manteniendo los márgenes brutos de los productores.

El Problema

Cuando los colaboradores van al campo a estudiar/monitorear las condiciones del mismo, ellos suben la información a una plataforma interna de la empresa.

La plataforma permite escribir una anotación, utilizar un mensaje de voz y agregar un *tag o etiqueta*.

Lo que termina sucediendo es que los colaboradores escriben todo lo que ven en una de las anotaciones. Solo el 5% le asigna un *tag*. Además esas anotaciones quedan como texto plano en la plataforma sin agregar ningún tipo de valor cuantificable

En el campo, las condiciones (sol, calor, lloviznas, etc) no son favorables para utilizar mucho tiempo en asignar un tag.

Cómo procesar esas anotaciones para darle valor?

Las Anotaciones

“Se observa presencia de rama negra”

“Cruciferas 20% cobertura”

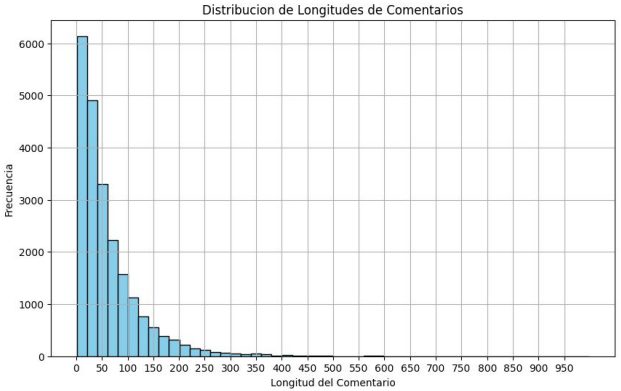
“Buen control de nabo”

“Sorgo de.alepo manchones. Abrojos x manchones”

“Se aplicaron 2 lts de sulfosato mas 1 lt de cletodim Transpect al 24% el 21 de Abri”

“Campo limpio, no hay presencia de rama negra”

- Ventaja a explotar: utilización de términos tecnicos específicos (ej: crucíferas, alepo, roya)
- Potencial problema a tratar: dificultad para eliminar las stop words sin análisis previo (ej: “no se observa presencia de ...”, la palabra “no” es fundamental)
- Debido al apuro del colaborador, las anotaciones suelen ser cortas (generalmente no más de 100 caracteres)



Atención a las Stop Words

Se observa que hay palabras combinadas como “cultivo con”, “sin inconveniente”, “con presencia” las cuales nos obligan a tener un especial cuidado con las stop words que decidimos eliminar.

Optamos por dejar dentro de las anotaciones las palabras:

- No, ni
- Sin, pero, aunque
- Muy, poco, mucho
- Con, sobre, hay

La Propuesta

Nos proponemos clasificar un porcentaje medio (en torno al 50%), en diversas categorías que surgen de diccionarios de productos e indicadores. Por ejemplo:

- Fungicidas
- Herbicidas
- Insecticidas
- Enfermedades
- Malezas
- Plagas

No nos interesan anotaciones puntuales, sino observar tendencias

No necesitamos clasificar el 100% de ellas



Diccionario de Términos Técnicos

En nuestra situación particular:

Aprovechamos el uso de **palabras técnicas** como marcas de fungicidas, herbicidas o insecticidas, y también nombres comunes y científicos de enfermedades, malezas o plagas.

Ayuda mucho en la clasificación de anotaciones

Confeccionamos un diccionario de palabras técnicas

Con base en un archivo excel que la empresa tiene armado, asociado a las 6 categorías de clasificación.

Si el comentario tiene una de esas palabras, se *tagea* como la categoría correspondiente.

Bigrama y Ejemplo de Tageo

Bigram	Tag
pasto cuaresma	malezas
soja guacha	malezas
roya anaranjada	enfermedades
mancha amarilla	enfermedades
rama negra	malezas

Bigrama de comentarios

Para este ejemplo vemos cómo del diccionario podemos extraer tokens con su tag específico utilizando las bases de la empresa.

Por otro lado tenemos el bigrama extraído de los comentarios con sus frecuencias. El objetivo es utilizar diferentes medidas (Coseno, Euclidea y Levenshtein) para lograr buscar similitudes cercanas entre ambas y así lograr un join entre diccionario y bigrama de comentarios con el tag definitivo sobre estos últimos.

Métodos para el Cálculo de Distancias/Similitudes

Característica	Coseno	Euclidea	Levenshtein
Tipo	Similitud	Distancia	Distancia
Rango	[-1, 1] o [0, 1]	[0, ∞)	[0, max(len(s1), len(s2))]
Uso principal	Comparación de vectores, documentos	Medición de distancia en espacios n-dimensionales	Comparación de cadenas de texto
Sensibilidad a la magnitud	No	Sí	No aplica
Ventajas	- Independiente de la magnitud - Eficaz para espacios de alta dimensión	- Intuitiva - Fácil de calcular - Útil en espacios de baja dimensión	- Útil para corrección ortográfica - Mide similitud de cadenas
Desventajas	- No considera la magnitud de los vectores	- Sensible a outliers - Puede ser menos útil en alta dimensión	- Computacionalmente costosa para cadenas largas

Tratamiento de palabras cortas → Dividimos la distancia por *len(word)* → Así evitamos que las palabras cortas generen valores de distancias chicas siempre.

Resultados

Bigram	Phrase	Tag	Score
karate zeon	insecticida	karate zeon	0.00
royas anaranjada	enfermedades	roya anaranjada	0.06
choris virgata	malezas	chloris virgata	0.07
rama negras	malezas	rama negra	0.09
mami guacho	malezas	mani guacho	0.09
rama negrs	malezas	rama negra	0.10
lecheron chamico	malezas	lecheron chico	0.12
bolsita pastor	malezas	bolsa pastor	0.14
pastoreo colorado	malezas	pasto colorado	0.18
maiz wacho	malezas	maiz guacho	0.20
x borreria	malezas	borreria	0.20
seca mani	malezas	sacha mani	0.22
commelina morenita	malezas	commelina erecta	0.22
dula gold	herbicida	dual gold	0.22
enfermedad ve	enfermedades	enfermedades	0.23

Distancias de Levenshtein

Con el uso de la distancia de Levenshtein, logramos asignar más del 40% de las etiquetas al establecer un punto de corte menor a 0.4.

Las instancias que superan este umbral fueron clasificadas como "sin etiquetar".

best_tag	
sin etiquetar	0.57
malezas	0.22
general lote	0.05
plagas	0.04
enfermedades	0.03
fungicida	0.02

Proporción de etiquetas generadas

Conclusiones

La aplicación de técnicas de Text Mining y el uso de la distancia de Levenshtein permite etiquetar más del 40% de los comentarios, para posteriormente:

1. Asociar comentarios a contextos temporales y geográficos.
2. Identificar temas técnicos específicos por región y tiempo.
3. Facilitar estrategias de marketing focalizadas.
4. Crear "mapas de calor" para visualizar la distribución de temas específicos.

Esta mejora en la clasificación proporciona insights valiosos para una toma de decisiones informadas y una respuesta efectiva a las necesidades regionales, optimizando las estrategias de marketing y atención al cliente.

Para optimizar los resultados en la correcta asignación de etiquetas, sería recomendable entrenar clasificadores más avanzados, que puedan superar el desempeño de modelos como NaiveBayesClassifier, MaxentClassifier y SklearnClassifier con MultinomialNB.