In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

In [2]:

```
data = pd.read_csv("humour_data.csv")
```

In [3]:

```
data.head()
```

Out[3]:

	text	humor
0	Joe biden rules out 2020 bid: 'guys, i'm not r...	False
1	Watch: darvish gave hitter whiplash with slow ...	False
2	What do you call a turtle without its shell? d...	True
3	5 reasons the 2016 election feels so personal	False
4	Pasco police shot mexican migrant from behind,...	False

In [4]:

```
data.tail()
```

Out[4]:

	text	humor
199995	Conor maynard seamlessly fits old-school r&b h...	False
199996	How to you make holy water? you boil the hell ...	True
199997	How many optometrists does it take to screw in...	True
199998	Mcdonald's will officially kick off all-day br...	False
199999	An irish man walks on the street and ignores a...	True

In [5]:

```
data.shape
```

Out[5]:

```
(200000, 2)
```

In [6]:

```
data.columns
```

Out[6]:

```
Index(['text', 'humor'], dtype='object')
```

In [7]:

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200000 entries, 0 to 199999
Data columns (total 2 columns):
 #   Column  Non-Null Count   Dtype 
---  ------  --------------   ----- 
 0   text    200000 non-null  object
 1   humor   200000 non-null  bool  
dtypes: bool(1), object(1)
memory usage: 1.7+ MB
```

In [9]:

```
data.isnull().sum()
```

Out[9]:

```
text     0
humor    0
dtype: int64
```

In [10]:

```
data['humor'].value_counts()
```

Out[10]:

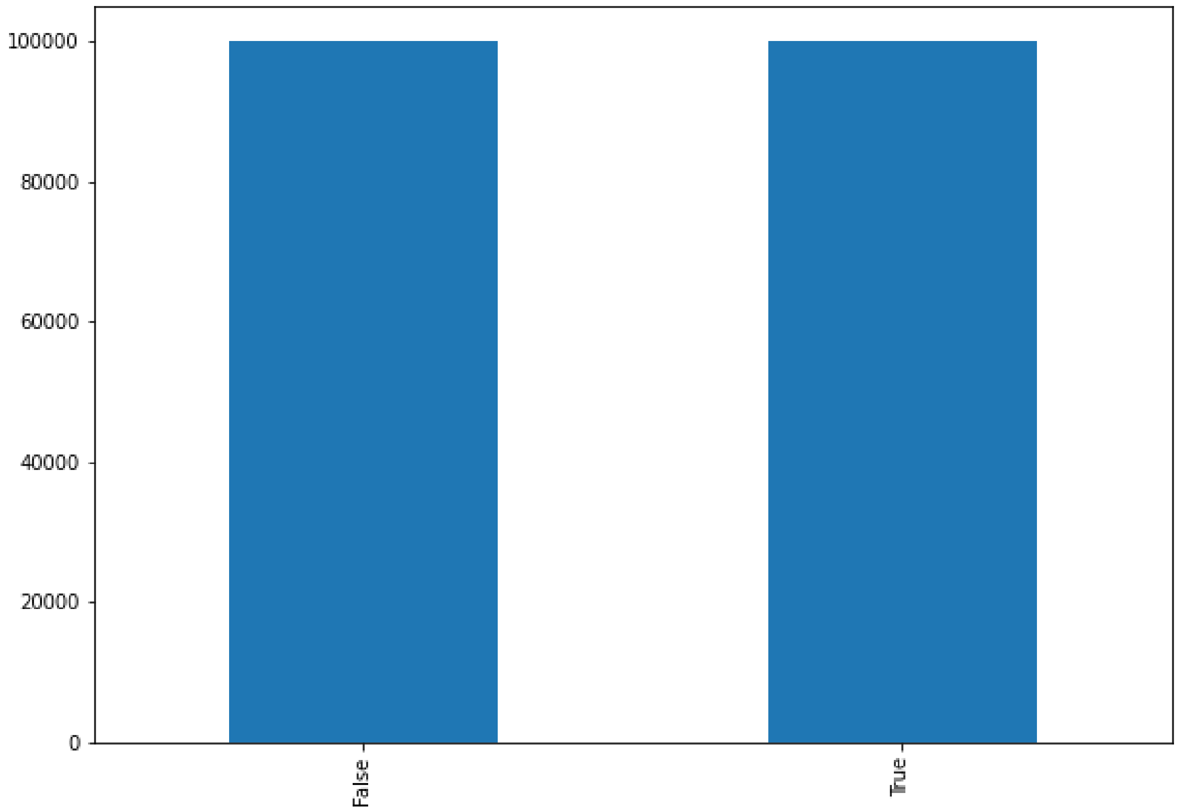
```
False    100000
True     100000
Name: humor, dtype: int64
```

In [11]:

```
data1 = data['humor'].value_counts()
```

In [18]:

```
fig = plt.figure(figsize =(10, 7))
data1.plot.bar()
plt.show()
```



In [19]:

```
humor_data = data[data['humor'] == True]
humor_data
```

Out[19]:

	text	humor
2	What do you call a turtle without its shell? d...	True
6	What is a pokemon master's favorite kind of pa...	True
7	Why do native americans hate it when it rains ...	True
9	My family tree is a cactus, we're all pricks.	True
13	How are music and candy similar? we throw away...	True
...	...	...
199990	Where do eskimos keep their money? in snowbanks.	True
199993	What did the child with no arms get for christ...	True
199996	How to you make holy water? you boil the hell ...	True
199997	How many optometrists does it take to screw in...	True
199999	An irish man walks on the street and ignores a...	True

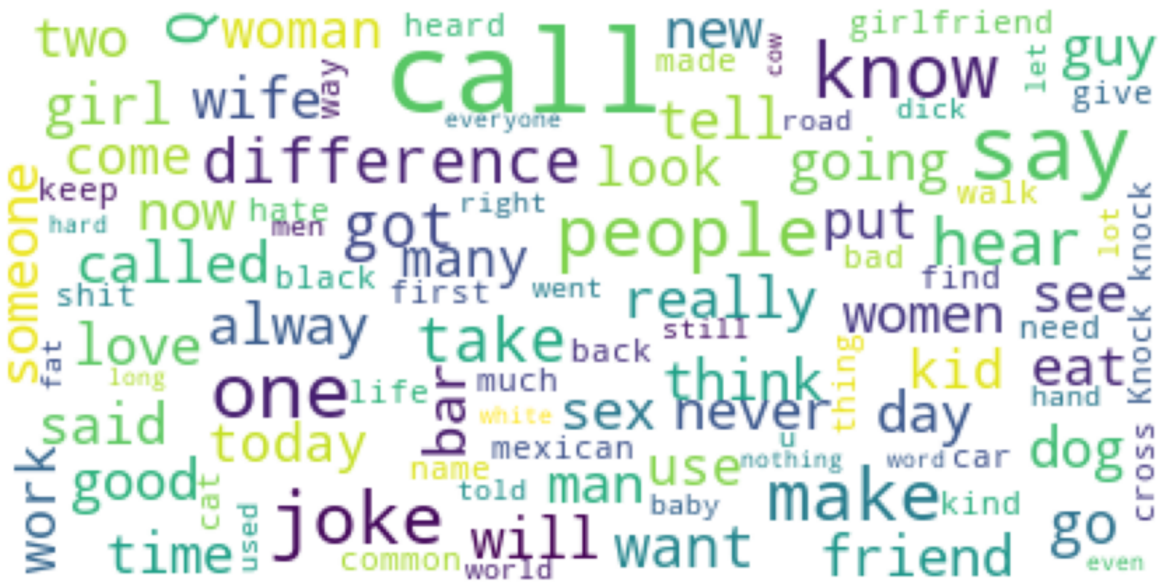
100000 rows × 2 columns

In [20]:

```
from wordcloud import WordCloud, STOPWORDS
texts = ' '.join(humor_data['text'])
stopwords = STOPWORDS

wordcloud = WordCloud(background_color='white',
                      stopwords=stopwords,
                      max_words=100,
                      max_font_size=40,
                      random_state=42).generate(texts)

plt.figure(figsize = (15, 12), facecolor = None)
plt.axis('off')
plt.imshow(wordcloud);
```



In [21]:

```
non_humor_data = data[data['humor'] == False]
non_humor_data
```

Out[21]:

	text	humor
0	Joe biden rules out 2020 bid: 'guys, i'm not r...	False
1	Watch: darvish gave hitter whiplash with slow ...	False
3	5 reasons the 2016 election feels so personal	False
4	Pasco police shot mexican migrant from behind,...	False
5	Martha stewart tweets hideous food photo, twit...	False
...	...	...
199991	Meet the billionaire who controls your ketchup...	False
199992	North korea stages large-scale artillery drill...	False
199994	Elizabeth taylor looked amazing even without d...	False
199995	Conor maynard seamlessly fits old-school r&b h...	False
199998	Mcdonald's will officially kick off all-day br...	False

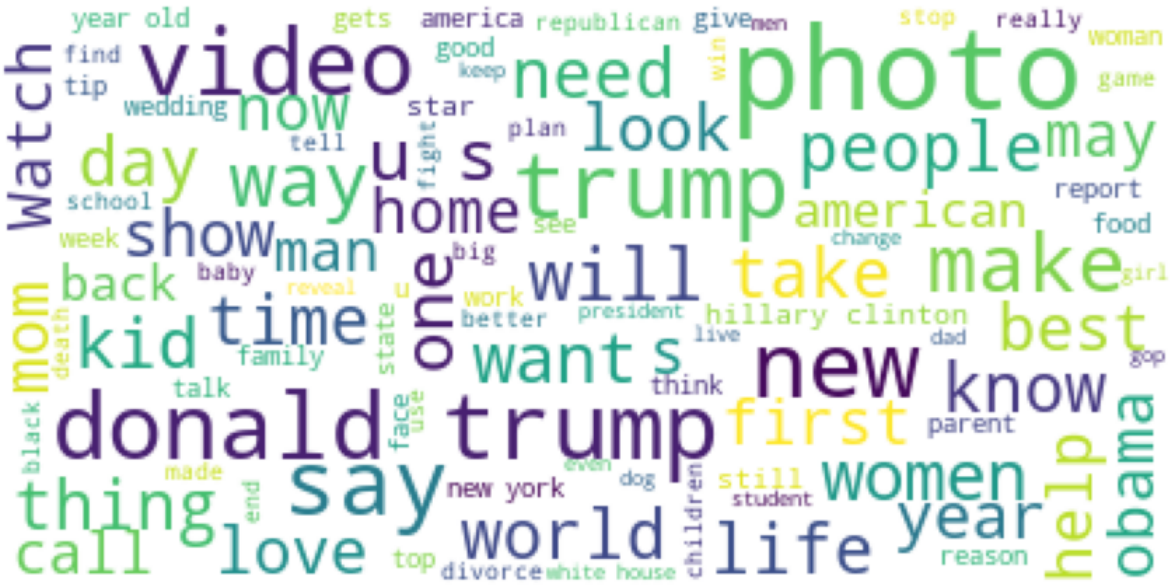
100000 rows × 2 columns

In [22]:

```
texts = ' '.join(non_humor_data['text'])
stopwords = STOPWORDS

wordcloud = WordCloud(background_color='white',
                      stopwords=stopwords,
                      max_words=100,
                      max_font_size=40,
                      random_state=42).generate(texts)

plt.figure(figsize = (15, 12), facecolor = None)
plt.axis('off')
plt.imshow(wordcloud);
```



In [23]:

```
data['question'] = data['text'].str.contains('\?')
data
```

Out[23]:

	text	humor	question
0	Joe biden rules out 2020 bid: 'guys, i'm not r...	False	False
1	Watch: darvish gave hitter whiplash with slow ...	False	False
2	What do you call a turtle without its shell? d...	True	True
3	5 reasons the 2016 election feels so personal	False	False
4	Pasco police shot mexican migrant from behind,...	False	False
...	...	...	...
199995	Conor maynard seamlessly fits old-school r&b h...	False	False
199996	How to you make holy water? you boil the hell ...	True	True
199997	How many optometrists does it take to screw in...	True	True
199998	Mcdonald's will officially kick off all-day br...	False	False
199999	An irish man walks on the street and ignores a...	True	False

200000 rows × 3 columns

In [24]:

```
data.groupby(['question', 'humor']).count()
```

Out[24]:

		text
question	**humor**	
False	**False**	94745
	True	46944
True	**False**	5255
	True	53056

In [28]:

```
import re
import string
import nltk
from nltk.util import pr
from nltk.corpus import stopwords
stemmer = nltk.SnowballStemmer("english")
nltk.download('stopwords')
stopword=set(stopwords.words('english'))
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\pc\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

In [29]:

```
def clean(text):
    text = str(text).lower()
    text = re.sub('\[.*?\]', '', text)
    text = re.sub('https?://\S+|www\.\S+', '', text)
    text = re.sub('<.*?>+', '', text)
    text = re.sub('[%s]' % re.escape(string.punctuation), '', text)
    text = re.sub('\n', '', text)
    text = re.sub('\w*\d\w*', '', text)
    text = [word for word in text.split(' ') if word not in stopword]
    text=" ".join(text)
    text = [stemmer.stem(word) for word in text.split(' ')]
    text=" ".join(text)
    return text
data["text"] = data["text"].apply(clean)
```

In [30]:

```
from sklearn.tree import DecisionTreeClassifier
```

In [31]:

```
x = np.array(data["text"])
y = np.array(data["humor"])
```

In [32]:

```
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.model_selection import train_test_split
```

In [33]:

```
cv = CountVectorizer()
X = cv.fit_transform(x)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=
```

In [34]:

```
clf = DecisionTreeClassifier()
clf.fit(X_train,y_train)
```

Out[34]:

```
DecisionTreeClassifier()
```

In [35]:

```
text1 = "I like sleeping. I just dont like going to sleep."
data = cv.transform([text1]).toarray()
print(clf.predict(data))
```

```
[ True]
```

In [36]:

```
text2 = "I wonder if the earth, teases other planets, for having no life."
data = cv.transform([text2]).toarray()
print(clf.predict(data))
```

```
[False]
```

In [37]:

```
text3 = "In the morning, there is a huge difference between 6:00 and 6:10."
data = cv.transform([text3]).toarray()
print(clf.predict(data))
```

```
[ True]
```

In [38]:

```
print("Training Accuracy :", clf.score(X_train, y_train))
print("Testing Accuracy :", clf.score(X_test, y_test))
```

```
Training Accuracy : 1.0
Testing Accuracy : 0.8246818181818182
```